

Proceedings X-Meeting 2016

Editor: AB³C

January 5, 2017

Conference Program

1 Organizing Committee	1
2 Introduction	3
3 Abstracts	5

Poster Session	5
-----------------------	----------

Genes and Genomics	5
---------------------------	----------

- 6 Design of chimeric antigens of Porcine Reproductive and Respiratory Syndrome Virus (PRRSV) through bioinformatics approaches: a rational model for the development of a diagnostic test
Jerusa Botelho Souza, Giuliana Loreto Saraiva, Jackson de Andrade Teixeira, Pedro Marcus Pereira Vidigal, Márcia Rogéria de Almeida Lamêgo
- 7 Prediction of microRNAs and miRNA pathway genes in Solanum lycopersicum and Solanum pennellii
Thaís Cunha de Sousa Cardoso, Tamires Caixeta Alves, Carolina Milagres Caneschi, Douglas Santana, Laurence Rodrigues do Amaral, Luiz Antônio Augusto Gomes, Wilson Roberto Maluf, Matheus de Souza Gomes
- 8 In silico genomic analysis of the endophytic bacterium Bacillus amyloliquefaciens 629
Brena Mota Moitinho Sant'Anna, Artur Trancoso Lopo de Queiroz, Milton Ricardo de Abreu Roque
- 9 Genotoxicity testing in-silico: quantification of the DNA mutation caused by the glycosidic bond hydrolysis
Bárbara Zanandreiz de Siqueira Mattos, Gustavo Henrique Passini Santos, Anton Semenchenko
- 10 Admixture Mapping of Brazilians Identifies New Obesity Susceptibility Loci
Hanaísa de Pla e Sant Anna, Marilia Sciliar, Meddly L. Santolalla Robles, Thiago Peixoto Leal, Gilderlanio Santana de Araújo, Mateus Gouveia, Wagner Magalhaes, Fernanda Kehdy, Eduardo Martin Tarazona Santos
- 11 Diagnostic metagenomics: a case study based on suspected dengue infection
Liliane Conteville, Michel Abanto Marín, Ana Maria Bispo de Filippis, Rita Maria Ribeiro Nogueira, Marcos César Lima de Mendonça, Ana Carolina Paulo Vicente
- 12 Detection of Functional Analogous Enzymes in the Human Metabolism
Rafael Mina Piergiorgi, Ana Carolina Ramos Guimarães, Marcos Paulo Catanho de Souza
- 13 tRNA array genomic survey unravels their presence and structure in Mycobacteria
Sergio Mascarenhas Morgado, Ana Carolina Paulo Vicente, Michel Abanto Marín
- 14 The first complete genome sequence of Streptococcus dysgalactiae subsp. dysgalactiae an emerging fish pathogen
Alexandra Antonieta Urrutia Zegarra, Felipe Luiz Pereira, Fernanda Alves Dorella, Alex F. Carvalho, Gustavo Morais Barony, Carlos A. G. Leal, Henrique Figueiredo
- 15 Assembly, annotation and comparison of Corynebacterium pseudotuberculosis lineages
Doglas Parise, Thiago de Jesus Sousa, Mariana Teixeira Dornelles Parise, Adrian Valentín Muñoz Bucio, Felipe Luiz Pereira, Fernanda Alves Dorella, Efrén Díaz Aparicio, Henrique Figueiredo, Daniela Arruda Costa, Vasco A de C Azevedo
- 16 Identification of Specific Enzymes in the Comparison between Fusarium oxysporum and Arabidopsis thaliana
Larissa Catharina Costa, Nicolas Carels
- 17 A comparative in silico linear B-cell epitope prediction for South American and African Trypanosoma vivax strains
Rafael Lucas Muniz Guedes, Carla Monadeli Filgueira Rodrigues, Luiz Gonzaga Paula de Almeida, Paola Mino-pri, Marta Maria Geraldes Teixeira, Ana Tereza Ribeiro de Vasconcelos
- 18 Within and between gene variants: tracking for potential targets for populational linkage according to the metagenomic profile in a changing freshwater environment
Marcele Laux, Ricardo Assunção Vialle, José Miguel Ortega, Alessandra Giani

- 19 Draft genome of *Serratia marcescens* UENF 22-GI: a plant growth promoting bacterium isolated from vermicompost
Filipe Pereira Matteoli, Pollyanna Santiago Lopes, Fábio Lopes Olivares, Thiago Motta Venancio
- 20 Variant in the PDE4B related to Acute Lymphoblastic Leukemia relapse is differentiated in Native Americans
Rennan Garcias Moreira, Fernanda Rodrigues Soares, Eduardo Martin Tarazona Santos
- 21 LTR retrotransposons in *Hemileia vastatrix* genome
Rafaela Leite Prado Rocha, Pedro Ricardo Marques Barreiros, Tiago Antônio de Oliveira Mendes, Laércio Zambolim, Ney Sussumu Sakiyama, Eveline Teixeira Caixeta
- 22 Linear Algebra Methods for Inferring Phylogenies Based on Peptides Frequencies Vectors: An Efficient Alternative Method to Investigate Relationships among Genes, Genomes and Organisms
Braulio Roberto Gonçalves Marinho Couto, Lara Maria Silva Miranda, Gabriel Bandeira Tofani, Gustavo Palmer Irffi, Lucas Felipe Silva, Matheus Allef Cruz, Thiago do Carmo Libreton Rocha, Marcos Augusto dos Santos
- 23 The genomic basis for the variable biochemical profiles that lead to erroneous identifications of emerging pathogenic *Corynebacterium* spp.
André de Souza Santos, Luis Gustavo Carvalho Pacheco, Carolina S. Silva, Catarina A. Moreira, Artur Silva, Vasco A de C Azevedo, Liza F. Vilela
- 24 Inferring the genetic structure and the history of interaction between Andean and Amazonian human populations using genome-wide data
Victor Octavio Borda Pua, Marilia Scliar, Mateus Gouveia, Thiago Peixoto Leal, Gilderlanio Santana de Araújo, Giordano Souza, Robert H Gilman, Heinrich Guio, Eduardo Martin Tarazona Santos
- 25 Genotypic characterization of *Vibrio parahaemolyticus* strains isolated in Brazil
Cristóvão Antunes de Lanna, Leandro de Oliveira Santos, Paulo Bisch, Wanda von Krieger
- 26 Searching for genomic elements of sexual reproduction in a microsporidian pathogen
Juliano De Oliveira Silveira, Karen Luisa Haag, Jean-François Pombert
- 27 Identification and variability analysis of monooxygenase gene family from *Chrysoporthe cubensis*
Túlio Morgan, Murillo Peterlini Tavares, Rafaela Inês de Souza Ladeira Ázar, Tiago Antônio de Oliveira Mendes, Valéria Monteze Guimarães
- 28 In silico prediction of auxiliary activity enzymes secreted by the fungus *Chrysoporthe cubensis*
Murillo Peterlini Tavares, Túlio Morgan, Thiago Rodrigues Dutra, Tiago Antônio de Oliveira Mendes, Hugo Rody Vianna Silva, Valéria Monteze Guimarães
- 29 A novel hierarchical in silico approach for the prediction of drug and vaccine targets against *Chlamydophila pneumoniae*
Ana Carolina Barbosa Caetano,
- 30 Establishment of a pipeline for 16S-based metagenomic studies of *Mycobacterium leprae*
Felipe Borim Correa, Gabriel da Rocha Fernandes
- 31 Impact of non-synonymous mutations in adaptive diversification and domestication of soybean
Kanhu charan Moharana, Thiago Motta Venancio
- 32 Prediction and analysis of plasmids from multidrug-resistant *Klebsiella pneumoniae* and *Enterobacter aerogenes* clinical isolates
Hemanuel Passarelli Araujo, Filipe Pereira Matteoli, Jussara Kasuko Palmeiro, Líbera Dalla-Costa, Thiago Motta Venancio
- 33 Chromosomal copy number variation reveals extensive levels of genomic plasticity among and within *Trypanosoma cruzi* DTUs
João Luís Reis Cunha, Gabriela Flavia Rodrigues Luiz, Hugo O. Valdivia, Rodrigo P. Baptista, Laila Almeida, Mariana Santos Cardoso, Tiago Antônio de Oliveira Mendes, Andrea M. Macedo, Ana Tereza Ribeiro de Vasconcelos, Gustavo Coutinho Cerqueira, Daniella Bartholomeu

- 34** Modular variability of multigene families encoding surface proteins uncovers differential composition of motifs among *Trypanosoma cruzi* strains
João Luís Reis Cunha, Gabriela Flavia Rodrigues Luiz, Rodrigo P. Baptista, Laila Almeida, Mariana Santos Cardoso, Gustavo Coutinho Cerqueira, Daniella Bartholomeu
- 35** Genomic identification and patterns of expression of secondary metabolite gene clusters in the entomopathogen fungus *Metarhizium anisopliae*.
Augusto Schrank, Nicolau Sbaraini
- 36** Genome analysis of *E. nigrum* and other filamentous fungi reveals molecular mechanisms related to endophytic/pathogenic lifestyles
Almir José Ferreira, Liliane Santana Oliveira Kashiwabara, João M. P. Alves, Michael Thon, Alan M. Durham, Léia C. L. Fávaro, Arthur Gruber, Wellington L. Araújo
- 37** Combining profile Hidden Markov Models and small RNA pattern based strategies to identify novel Endogenous Viral Elements (EVEs) and exogenous viruses
Arthur Gruber, Eric Roberto Guimarães Rocha Aguiar, Liliane Santana Oliveira Kashiwabara, João M. P. Alves, João T. Marques
- 38** Preliminary analysis of functional SNPs mining from GBS data using GATK in a sugarcane map population
Alexandre Hild Aono, Estela Araujo Costa, Hugo Rody Vianna Silva, James Shiniti Nagai, Anete Pereira de Souza, Reginaldo Massanobu Kuroshu
- 39** GBS data from a population of modern sugarcane variety reveals duplicate genes retained by breeding process
James Shiniti Nagai, Hugo Rody Vianna Silva, Estela Araujo Costa, Alexandre Hild Aono, Anete Pereira de Souza, Reginaldo Massanobu Kuroshu
- 40** Characterization of phage sequences on *Corynebacterium pseudotuberculosis* genomes
Flavia Figueira Aburjaile, Amália Raiana F. Lobato, Luís Carlos Guimarães, Ana Lídia Queiroz Cavalcante, Kenny da Costa Pinheiro, Adonney Allan de Oliveira Veras, Rafael Azevedo Baraúna, Artur Silva, Rommel Thiago Jucá Ramos
- 41** Comparative genomic analysis of clinical and environmental strains of *Vibrio parahaemolyticus* isolated in Brazil: insight into their virulence potential
Leandro de Oliveira Santos, Paulo Bisch, Wanda von Krüger
- 42** Identification of Pho regulon genes and Pho box-like sequences in genomes of clinical and environmental isolates of *Vibrio parahaemolyticus* from Brazil
Leandro de Oliveira Santos, Cristóvão Antunes de Lanna, Paulo Bisch, Wanda von Krüger
- 43** Identification of genetic variations in engineered yeast for xylose consumption and acetic acid resistance applied to second generation ethanol production
Sheila Tiemi Nagamatsu, Luige Armando Llerena Calderon, Lucas Parreiras, Bruna Tatsue, Angelica Martins Gomes, Gonçalo Amarante Guimarães Pereira, Marcelo Falsarella Carazzolle
- 44** In silico approaches to predict the impact of leukemic DNMT3a mutations & Identification of leads based on drug decitabine using Complex Based Pharmacophore Mapping and Virtual Screening
Syed babar Jamal Bacha, Matheus Filgueira Bezerra, sandeep tiwari, Flavia Figueira Aburjaile, Vasco A de C Azevedo, Artur Silva, Cintia Renata Rocha Costa, Marcos André Calvancanti Bezerra, Antonio Roberto Lucena-Araujo, Eduardo Isidoro Carneiro Beltrão
- 45** In-silico analyses for the discovery of drug and vaccine targets in *Corynebacterium camporealensis*: A Novel Hierarchical Approach
Syed babar Jamal Bacha, sandeep tiwari, Arun Kumar Jaiswal, Daniela Arruda Costa, USUARIO TESTE, Doglas Parise, Henrique CP Figueiredo, Debmalya Barh, Artur Silva, Vasco A de C Azevedo
- 46** Genotype imputation of Hereford and Bradford bovine breeds from Brazil
MAURICIO DE ALVARENGA MUDADU, Henry Gomes de Carvalho, Marcos Jun Iti Yokoo, Fernando Flores Cardoso
- 47** Detection of potential genetic variants affecting gene function in Guzerat cattle
Adhemar Zerlotini Neto,

- 48 Characterization of the probiotic and stress resistance-related genes of *Lactococcus lactis* subsp. *lactis* NCDO 2118 through comparative genomics and in vitro assays.
Letícia de Castro Oliveira,
- 49 A comparison of two pipelines for metagenomic 16S rRNA using Ion Torrent (PGM) Sequencing Platform
Daniel Vasconcelos Rissi,Suzana Eiko Sato Guima,Rodrigo Matheus Pereira
- 50 Insights into *Klebsiella pneumoniae* type VI secretion system regulation
Victor Barbosa,Leticia MS Lery
- 51 AMP-Identifier: A Unix shell script for antimicrobial peptide identification
João Pacifico Bezerra Neto,Mauro Guida dos Santos,Ana Maria Benko-Iseppon
- 52 Genome mining of biosynthetic gene clusters in *Nostoc* sp. CACIAM 19, a cyanobacterium from an Amazonian environment
DAVID BATISTA MAUÉS,ALEX RANIERI JERÔNIMO LIMA,PABLO HENRIQUE GONÇALVES MORAES,Andrei Santos Siqueira,Leonardo Teixeira Dall'Agnol,EVONNILDO COSTA GONÇALVES
- 53 Ab initio characterization of promoter regions based on Conditional Random Fields
Ígor Bonadio,Mauro de Medeiros Oliveira,Alan Durham
- 54 Study of Chromatin Remodeling in Colorectal Cancer Progression
Simone Nantes de Aquino,Nicole Scherer,Mariana Boroni
- 55 Taxonomic identification of metagenomics reads based on sequence features by Support Vector Machines (SVM)
Tahila Andriguetti,Ney Lemke
- 56 Cell cycle and metabolism related candidate human synthetic lethal network
sandeep tiwari,Thiago Luiz de Paula Castro,Núbia Seiffert,Debmalya Barh,Vasco A de C Azevedo
- 57 The discovery of novel multiple small deletions within human coding genes associated to known lung cancer pathways.
Gabriel Wajnberg,Raphael Tavares da Silva,Nicole Scherer,Carlos Gil Ferreira,Fabio Passetti
- 58 Identification of somatic mutations in prostate adenocarcinoma with Gleason score 7 and 8 and their associations with biochemical recurrence
Isabella Tanus Job e Meira,Bruna D F Barros,Rodrigo F Ramalho,José E Kroll,Renan Valieres,Sandro Jose de Souza,Isabela W da Cunha,Gustavo C Guimarães,Dirce M Carraro,Elisa N Ferreira
- 59 Detection and correction mis-assemblies in genome of *Corynebacterium pseudotuberculosis*
Thiago de Jesus Sousa,Doglas Parise,Diego César Batista Mariano,Daniela Arruda Costa,Felipe Luiz Pereira,Henrique Figueiredo,Artur Silva,Rommel Thiago Jucá Ramos,Vasco A de C Azevedo
- 60 Genomic analysis of opportunistic bacteria from *Herbaspirillum* genus isolated from immunocompromised patients
Helisson Faoro,Willian Klassen de Oliveira,Michelle Zibetti Tadra-Sfeir,Rodrigo Luis Cardoso,Emanuel Maltempi de Souza,Fabio de Oliveira Pedrosa
- 61 Investigation of mutations in the HBB gene using the 1000 GENOMES databank
Tania Carlice Lopes Pereira dos Reis,Jaime Viana,Fabiano Moreira Cordeiro,Greice de Lemos Cardoso,João Guerreiro,Sidney Santos,ÂNDREA KELY CAMPOS RIBEIRO DOS SANTOS
- 62 An hierarchical classification system for beta-lactamases
Melise Chaves Silveira,Fábio Mota,Rodrigo Jardim,Rangeline Azevedo da Silva,Marcos Paulo Catano de Souza,Ana Carolina Ramos Guimarães,antonio Basilio de Miranda
- 63 Draft genome sequence of the extremophile endemic marine antarctic yeast *Metchnikowia australis*
Heron Hilário,Thiago Mafra Batista,Rennan Garcias Moreira,Valéria Martins Godinho,Carlos Augusto Rosa,Luiz Henrique Rosa,Glória Regina Franco
- 64 Diversity analysis of Howler monkey (*Alouatta* spp.) fecal microbiota
Raquel Riyuko de Almeida Franco,Layla Martins,João Batista,Andrew Thomaz,Julio Oliveira,Aline Maria da Silva,Joao Carlos Setubal

- 65 Spacial Organization of Genomes: Insights on coordinated regulation of Biological Pathways
Luís Henrique, José Miguel Ortega
- 66 HD-zip classification in *Vigna unguiculata* and comparative synteny with *Phaseolus vulgaris*
Bruna Pierreck Moura, Artemisa Nazaré Costa Borges, Caroline de Jesus Pires, Flávia Tadeu de Araújo, José Ribamar Costa Ferreira-Neto, Ana Christina Brasileiro-Vidal, Ana Maria Benko-Iseppon
- 67 CattleQTLdb analysis to increase understanding of the functions of milk proteins genes
Carolina Guimarães Ramos Matosinho, Izinara Rosse da Cruz, Pablo Augusto de Souza Fonseca, Juliana Assis, Francilson Silva de Oliveira, Flávio Marcos Gomes Araújo, Anna Christina de Matos Salim, Wagner Arbex, Marco Antônio Machado, Maria Gabriela Campolina Diniz Peixoto, Rui da Silva Verneque, Marta Fonseca Martins, Roney Santos Coimbra, Marcos Vinícius Gualberto Barbosa da Silva, Guilherme Oliveira, Maria Raquel Santos Carvalho
- 68 In silico identification of the effects of genetic variants in transcription factors recognition sites in regulatory regions of candidate genes for reproductive disorders in cattle
Luiza de Almeida Ferreira Diniz, Pablo Augusto de Souza Fonseca, Ana Emilia de Paiva, Fernanda Caroline Santos, Izinara Rosse da Cruz, Maria Raquel Santos Carvalho
- 69 Metagenomic analysis of the Southern Brazilian Atlantic Forest soil using next-generation sequencing technologies
Janyinne Stephanie de Oliveira Palheta, Michelle Zibetti Tadra-Sfeir, Emanuel Maltempi de Souza, Fabio de Oliveira Pedrosa, Helisson Faooro
- 70 Inference of distant homologs in Protozoa by pHMM–pHMM (profile Hidden Markov Model) comparison for the identification of superfamilies
Darueck Acácio Campos, Alberto M. R. Dávila, Rodrigo Jardim
- 71 Comparative Genomics between two different biovars of *Corynebacterium pseudotuberculosis* isolated in the same host
Rafael Cabus Gantois, Thiago de Jesus Sousa, Doglas Parise, Daniela Arruda Costa, Anne Cybelle Pinto Gomide, Henrique Figueiredo, Vasco A de C Azevedo
- 72 Relative Evaluation of NoSQL Databases For Manipulating Genotype Data
Arthur Lorenzi Almeida, Vinicius Schettino, Fernanda Nascimento Almeida, Wagner Arbex
- 73 Metagenomics insights reveals functional patterns among soil microbial communities of global biomes
Melline Fontes Noronha, Gileno Vieira Lacerda Junior, Jack A Gilbert, Valéria Maia de Oliveira
- 74 Complete genome sequence of *Corynebacterium pseudotuberculosis* 33
Marcus Vinicius Canário Viana, Doglas Parise, Thiago de Jesus Sousa, Leandro de Jesus Benevides, Diego César Batista Mariano, Flávia de Souza Rocha, Priscilla Bagano, Luís Carlos Guimarães, Felipe Luiz Pereira, Fernanda Alves Dorella, Rommel Ramos, Salah Abdel Karim Selim, Mohammad Salaheldan, Artur Silva, Alice Rebecca Wattam, Vasco A de C Azevedo

Phylogeny and Evolution

75

- 75 Evolution of heterotrophy: Genes needed for regulation of the acidity of pancreatic juice appeared recently in man evolution
Fenícia Brito, Carlos Alberto Xavier Gonçalves, José Miguel Ortega
- 76 Walking through old routes to reach new destinations: unraveling the origin of the mammary gland
Lissur Azevedo Orsine, Elisa Rennó Donnard Moreira, José Miguel Ortega
- 77 Mosquitoes Mobilome
Gabriel da Luz Wallau, Elverson Soares de Melo
- 78 B-cell epitopes prediction in trypanosomatids genome core
Anderson Coqueiro dos Santos, Leandro Martins de Freitas
- 79 Construction of alternative algorithm for development of phylogenetic trees using InterPro entries and frequency vectors of tri-peptide
Braulio Roberto Gonçalves Marinho Couto, Lucas Felipe Silva, Dalila Dominique Duarte Rocha, Lara Maria Silva Miranda, Matheus Allef Cruz, Thiago do Carmo Librelon Rocha, Marcos Augusto dos Santos

- 80 Biological modules associated with prophage density in pathogenic and commensal Escherichia coli
Tarcisio José Domingos Coutinho,Francisco Pereira Lobo,Glória Regina Franco
- 81 Inferring the demographic history of Cnesterodon brevirostratus using bioinformatics tools
Daniela Ambros Quinsani,Martiela Vaz de Freitas,Aline M.C. Ramos-Fregonezi,Luis Roberto Malabarba,Nelson J.R. Fagundes
- 82 Improving the supertree approach by analyzing protein clusters with paralogs and including distance data
TETSU SAKAMOTO,José Miguel Ortega
- 83 Reconstructing ancestral protein-protein interactions of virus-host systems
Anderson Fernandes de Brito,John W. Pinney
- 84 Study of the origin of the genes controlling flower development
Beatriz Moura Kfouri de Castro,Lab Biodados,Carlos Alberto Xavier Gonçalves
- 85 Secondary structure changes according to evolutionary age
Ricardo Assunção Vialle,José Miguel Ortega
- 86 ProteinWorldDB in a multidimensional representation of the Network of Life
Edson Machado Filho,
- 87 GO-Genesis: finding the origin of biological processes and molecular functions from Gene Ontology
Carlos Alberto Xavier Gonçalves,Lab Biodados
- 88 mtDNA Data Mining: A Global Analysis
Camilla Reginatto De Pierri,Bruno Thiago de Lima Nichio,TETSU SAKAMOTO,Mauro Antônio Alves Castro,José Miguel Ortega,Roberto Tadeu Raitt
- 89 Tardigrades (Ecdysozoa: Tardigrada) A general review of the record in gene database of Cytochrome Oxydase I (COI) and the Damage Suppressor gene (DSUP).
Antonio A A Pires,
- 90 Highly resolved phylogeny for Corynebacteriales
NILSON DA ROCHA COIMBRA,Vasco A de C Azevedo,Aïda Ouangraoua
- 91 key amino acids in understanding evolutionary characterization of Mn/Fe-Superoxide Dismutase: A phylogenetic and structural analysis of proteins from Corynebacterium and hosts
Alberto F. de Oliveira Junior,Pammella Teixeira,Debmalya Barh,Preetam Ghosh,Vasco A de C Azevedo
- 92 Analysis of comparative genomics reveal evidence of positive selection on pathogenicity-related genes of Witches' broom disease on cocoa trees
Paulo Massanari Tokimatu Filho,Juliana José,Daniela Toledo Thomazella,Leandro Costa do Nascimento,Paulo Teixeira,Gonçalo Amarante Guimarães Pereira,Marcelo Falsarella Carazzolle
- 93 Analyzing molecular characteristics of small RNAs to assess the evolution of RNAi pathways
Eric Roberto Guimarães Rocha Aguiar,Flavia Viana Ferreira,Roenick Proveti Olmo,Karla Pollyanna Vieira de Oliveira,Simona Paro,Isaque João da Silva de Faria,Betânia Paiva Drumond,Vinicius Augusto Carvalho de Abreu,Carine Meignin,Maurício Roberto Viana Sant'Anna,Nelder de Figueiredo Gontijo,Luciano Andrade Moreira,Erna Geessien Kroon,Jean-Luc Imler,João T. Marques
- 94 Evaluation of the accuracy of the MLrelate kinship analysis program when no parents are sampled
LETICIA FERREIRA LIMA,Renata Schama
- 95 Assortative Mating in Brazilian Populations
isabela Alvim,Hanaisa de Pla e Sant Anna,Eduardo Martin Tarazona Santos
- 96 Insights into the population history of free-living bacteria as counted by their CRISPR inventory
Julliane Dutra Medeiros,Laura Rabelo Leite,Francislon Silva de Oliveira,Victor Satler Pylro,Gabriel da Rocha Fernandes,Guilherme Oliveira,Sara Cuadros Orellana
- 97 Evolutionary origin of the proteins involved in the entry of and defense to the Ebola virus in the host cell
ELISSON NOGUEIRA LOPES,TETSU SAKAMOTO,José Miguel Ortega
- 98 Study of new molecular markers for Phylogenetic reconstruction of the black fungus in humans
Edgar Lacerda de Aguiar,Cláudia Barbosa Assunção,Rachel Basques Caligorne

- 99 Molecular docking and structural optimization of bioactive compounds from natural products against 1-UAP of *L. braziliensis*
Sabrina Silva Mendonça, João Marcos Galúcio, Kelly Christina Ferreira Castro, Kauê Santana da Costa
- 100 Cluster analysis of AcrB protein molecular dynamics conformations
Núbia Souza Prates, Adriano Velasque Werhli, Karina Machado
- 101 Ligand-Based Pharmacophore Modeling and Virtual Screening of Plant-Derived Ligands for the Alpha-Amylase and Alpha-Glycosidase
Heitor Cappato, Nilson Nicolau Junior, Fouad Salmen Espindola
- 102 Structural Analysis of Alba Proteins of Leishmania infantum
Kauê Santana da Costa, Elvis Santos Leonardo, João Marcos Galúcio, Elcio Souza Leal, Jerônimo Lameira Silva
- 103 Non-Homology-Based Prediction of Protein Target Regions by Logistic Regression
Gustavo Santos de Oliveira, Marcos Augusto dos Santos, Vasco A de C Azevedo
- 104 Selecting structure-based virtual screening hits using chemoinformatics tools: a case study with HIV-1 reverse transcriptase
Lucianna Helene Silva dos Santos, Rafaela Ferreira, Ernesto Raul Caffarena
- 105 Characterization of the EF-IV (LepA) influence on mRNA translation by in-silico cell-free protein expression
Anton Semenchenko, Bárbara Zanandreiz de Siqueira Mattos, Guilherme Oliveira, A. P. F. Atman
- 106 GACP1 and HSP90.1 proteins may have an important role in the oxidative stress in Saccharum spp.
Felipe de Lima Almeida, Kellya Francisca Mendonça Barreto, João Paulo Matos Santos Lima, Katia Castanho Scortecci
- 107 Structural pattern detection for engineering more efficient enzymes for second-generation biofuel production
Diego Mariano, Thiago Silva Correia, José Renato Pereira de Moura Barroso, Raquel Melo Minardi
- 108 Using sequence weighting to improve residue correlation analysis
Lucas Carrijo de Oliveira, Lucas Bleicher
- 109 Mutational Analysis of the Virion Infectivity Factor (Vif) of HIV-1 subtype F2 and its influence on the interactions with APOBECs
Sabrina Silva Mendonça, João Marcos Galúcio, Kauê Santana da Costa, Elcio Souza Leal
- 110 Molecular interactions between NF-?B and thiopheneacetamide during mycobacteria infection
Vanessa dos Santos Silva, Ernesto Raul Caffarena, Fatima Vergara, Maria das Graças Henriques
- 111 Identification of druggable binding sites in ribose-5-phosphate isomerase of *T. cruzi*
Rafael Ferreira Soares, Ana Carolina Ramos Guimarães, Ernesto Raul Caffarena
- 112 VERMONT: A tool for mutation visualization
Sócrates Soares Araújo Júnior, Samuel da Silva Guimarães, Alexandre Victor Fassio, Raquel Melo Minardi, Sabrina de A. Silveira
- 113 New annotation strategies: from sequence to 3D structure
Rafael Nicolay B. da Silva, Paulo José Miranda da Silva Iwakami Beltrão, Manuela Leal da Silva
- 114 Metabolic changes of pathogenic and nonpathogenic Leishmania species during host cell infection by integration of mathematical models, quantitative proteomics and untargeted 1H-NMR
Tiago Antônio de Oliveira Mendes, Daniel Menezes Souza, Mariana Santos Cardoso, Alan Machado, Sebastião Rodrigo Ferreira, Laila Almeida, Alexandre Marques, Lúcia Pimenta, José Filho, Ernesto Nakayasu, Ricardo Fujiwara, Kiran Patil, Daniella Bartholomeu
- 115 Prospecting novel proteins from *Deinococcus radiodurans*: a model for putative heat shock proteins
Ricardo Valle Ladewig Zappala, Manuela Leal da Silva, Claudia A. S. Lage, Pedro Geraldo Pascutti
- 116 Definition and Comparative Analysis of the kinomes of *Leishmania infantum* and *L. braziliensis*.
Joyce Villa Verde Bastos Borba, Arthur de Carvalho e Silva, Pablo Ivan Pereira Ramos, Nicholas Furnham, Carolina Horta Andrade
- 117 Development of cruzain selective inhibitors by structure based virtual screening
Viviane Correa Santos, Rafaela Ferreira

- 118 Protein Folding by Generalized Simulated Annealing and Molecular Dynamics Methods
Pedro Geraldo Pascutti,Tácio Vinícius Amorim Fernandes
- 119 Mutation Analysis for AgrC from Staphylococcus aureus
Samuel da Silva Guimarães,Danielle Mendes Silva,Mônica Pacheco Silva,Pedro Marcus Vidigal,Andréa de Oliveira Barros Ribon,Sabrina de Azevedo Silveira
- 120 nAPOLI: a web tool for protein-ligand interactions analysis
Alexandre Victor Fassio,Sabrina de A. Silveira,Rafaela Ferreira,Raquel Melo Minardi
- 121 Characterization of an antimicrobial peptide from eggplant leaves as an inhibitor of carboxypeptidase
Maria Cristina Baracat Pereira,Victor Dose Lage de Almeida,Hebréia Oliveira Almeida-Souza,Maura Vianna Prates,Marcelo Porto Bemquerer,Tiago Antônio de Oliveira Mendes
- 122 Proteomic analysis of seminal plasma from stallions (Mangalarga Marchador) influenced by seasonality
Maria Cristina Baracat Pereira,Renato Lima Senra,João Gabriel da Silva Neves,Marcos Jorge Magalhaes Júnior,José Domingos Guimarães
- 123 GASS-WEB: a web server for identifying enzyme active sites based on genetic algorithms
João Pedro Areias de Moraes,Douglas E. V. Pires,Raquel Melo Minardi,Gisele L. Pappa,Sandro Izidoro
- 124 In silico structural studies of phospholipases A2 inhibitors from snake blood
Carlos A. H. Fernandes,Fábio F. Mattioli,Liza F. Vilela,Consuelo Latorre Fortes Dias,Marcos R. M. Fontes
- 125 In silico structural studies of Replication Proteins A1 and A2 from trypanossomatids (Leishmania and Trypanosoma)
Carlos A. H. Fernandes,Fábio F. Mattioli,Raphael S. Pavani,Marcos R. M. Fontes,Maria C. Elias,Maria I. N. Cano
- 126 Mutational Analysis of Human K-ras G12C and Design and Molecular Docking of ARS-853 derivatives
Kauê Santana da Costa,João Marcos Galúcio,Jose Cassio Figueiras
- 127 Modeling and Molecular Docking of the largest subunit of the Ribulose-1,5-Bisphosphate Carboxylase/Oxygenase (RuBisCO) from Alkalinema sp. CACIAM 70d.
James Siqueira Pereira,Andrei Santos Siqueira,Leonardo Teixeira Dall'Agnol,Juliana Simão Nina de Azevedo,Evonnildo C. Gonçalves
- 128 GReMLIN: A graph mining strategy to infer protein-ligand interaction patterns
Charles A. Santana,Fabio R. Cerqueira,Carlos H. da Silveira,Alexandre V. Fassio,Raquel Melo Minardi,Sabrina de A. Silveira
- 129 Modeling MS native-state amide hydrogen exchange through structural and dynamical properties.
Aline Beatriz Mello Rodrigues,Lucas de Almeida Machado,Luiz Max Fagundes de Carvalho,Mauricio Garcia Souza Costa,Leonardo Soares Bastos,Paulo Ricardo Batista
- 130 Comparison Between a Graph-Based Methodology and the LSQKAB to Check Proteins Similarities
Otaviano Martins Monteiro,Sandro Renato Dias,Thiago de Souza Rodrigues
- 131 Metabolic pathway prediction of enzymes: a machine learning approach
Rodrigo de Oliveira Almeida,Guilherme Targino Valente
- 132 In silico Identification of common putative vaccine candidates against Treponema pallidum: A reverse vaccinology based approach
Arun Kumar Jaiswal,sandeep tiwari,Syed babar Jamal Bacha,Vasco A de C Azevedo,Siomar de Castro Soares
- 133 T lymphocytes epitopes prediction to access immunological response from Chagas diseases patient samples
Cristiane Toledo,Júlia Castro,Ricardo T. Gazzinelli,Caroline Junqueira
- 134 Standardization of the ribosomal protein genes nomenclature in Leishmania major
Thaís Couto Laureano,
- 135 In silico prediction and comparative studies on plant GPCRs
Natalia Florencio Martins,Emmanuel Bresso,Priscila Grynberg,Roberto Coiti Togawa,Deisy X. Amora,Bernard Maigret

- 136 Bioinformatics approaches to identify, classify and prioritize protein kinases as drug targets in *Schistosoma japonicum*
Arthur de Carvalho e Silva, Joyce Villa Verde Bastos Borba, Pablo Ivan Pereira Ramos, Nicholas Furnham, Carolina Horta Andrade
- 137 Data Mining for characterization of nanotoxicity in mitochondrial ion channels induced by carbon nanotubes
Karina Machado, Luisa Rodrigues Cornetet, Michael Gonzalez Durruthy, Adriano Velasque Werhli, José Maria Monserrat
- 138 Prediction of protein stability changes upon single point mutation using Ensemble Learning
Adriano Velasque Werhli, Karina Machado, Alex D. Camargo
- 139 In Silico Binding Site Analysis of E6 Oncoproteins from High-Risk European HPV variants
Gabriel Monteiro da Silva, Elvira Tamarozzi, Silvana Giuliaatti
- 140 Analysis of Amino Acid coevolved sets in the Low Molecular Weight Phosphatase protein family by Molecular Dynamics
Marcelo Querino Lima Afonso, Lucas Bleicher
- 141 In silico intrinsic disorder analysis of β-crystallin B2 protein and mutations that cause congenital cataract
Júlia Barbieri de Oliveira Souza, José Eduardo Antônio Júnior, Elvira Tamarozzi, Silvana Giuliaatti
- 142 PPI-signature: homologous proteins interact with different partners in the same way
Larissa Fernandes Leijoto, Raquel Melo Minardi
- 143 Isocitrate Lyase of *Paracoccidioides brasiliensis*: Effects of Cofactor on Dynamical Stability and Virtual Screening of Natural Products
Luciane Sussuchi da Silva, Uessiley Ribeiro Barbosa, Fausto Guimarães Costa, Célia Maria de Almeida Soares, Maristela Pereira, Roosevelt Alves da Silva
- 144 Peptides modulators of malate synthase of *Paracoccidioides brasiliensis* obtained from Protein-Protein interactions and docking simulations
Roosevelt Alves da Silva, Raisa Melo Lima, Luciane Sussuchi da Silva, Gabriela Lima de Menezes, Célia Maria de Almeida Soares, Maristela Pereira
- 145 Inhibition Resistance Mechanism for the Product of Beta-Glucosidases, a Computational Approach
Rafael Eduardo Oliveira Rocha, Leonardo Henrique França de Lima
- 146 FlexSPS: A Monte Carlo update for protein refinement from I-TASSER models
Roosevelt Alves da Silva,
- 147 Identification of non-homologous isofunctional enzymes in the antioxidant system of plants and phytopathogens
Rangeline Azevedo da Silva, Leandro de Mattos Pereira, Melise Chaves Silveira, Monete Rajão Gomes, Ana Carolina Ramos Guimarães, Antonio Basilio de Miranda
- 148 In silico study of Hypoxanthine-guanine Phosphoribosyltransferase inhibitors for drug design against Leishmania species
Liliane Pereira de Araújo, Wagner Rodrigues de Assis Soares, Rosangela Santos Pereira, Bruno Silva Andrade
- 149 In silico screening of semi arid plant compounds targeting 5-Lipoxygenase (LOX)
Liliane Pereira de Araújo, Brenda Santana Portela, Wagner Rodrigues de Assis Soares, Rosangela Santos Pereira, Bruno Silva Andrade
- 150 Virtual screening of natural compounds from Brazilian semi arid plants targeting GABA receptor inhibitors
Wagner Rodrigues de Assis Soares, Gesivaldo Santos, Djalma Menezes de Oliveira, Vanderlúcia Fonseca de Paula, Bruno Silva Andrade
- 151 24-c-sterol-methyltransferase as a target for the design of new anti-trypanosomatids drugs
Manuela Leal da Silva, Kandy Anny de Azevedo Werneck, Gonzalo Guillermo Visbal Silva, Diego Enry Barreto Gomes
- 152 Isocitrate lyase protein-protein interaction assay of *Paracoccidioides* spp.
Kleber Santiago Freitas e Silva, Célia Maria de Almeida Soares, Maristela Pereira

- 153** In silico repurposing of approved drugs for paracoccidioidomycosis
Kleber Santiago Freitas e Silva, Amanda Alves de Oliveira, Bruno Junior Neves, Lívia do Carmo Silva, Célia Maria de Almeida Soares, Carolina H. Andrade, Maristela Pereira
- 154** Screening of compounds candidate to inhibit the interaction monomer - monomer of the NS1 protein of dengue virus: an approach for docking and molecular dynamics.
Ricardo Lemes Gonçalves, Luciane Sussuchi da Silva, Roosevelt Alves da Silva
- 155** Development of a peptide-based electrochemical biosensor for juvenile idiopathic arthritis diagnosis
Vinícius de Rezende Rodovalho, Galber Rodrigues Araujo, Carlos Ueira Vieira, João Marcos Madurro, Ana Graci Brito-Madurro
- 156** Automation of polyproteins GAG and GAG-POL-1 cleavage site mapping: A study of large scale molecular docking
Fernando Limoeiro Lara de Oliveira, Maria Fernanda Ribeiro Dias, Manuela Leal da Silva
- 157** Characterization of the proteome of four strains of *Lactococcus lactis* with biotechnological relevance
Caroline Leonel Vasconcelos de Campos, Wanderson Marques da Silva, Cassiana Severiano de Sousa, Siomar de Castro Soares, Guilherme Campos, Cristiana Perdigão Resende, Felipe Luiz Pereira, Gustavo Henrique Souza, Henrique Figueiredo, Vasco A de C Azevedo
- 158** Refining the calibration of a coarse-grained force field for protein complexation
Sergio Alejandro Poveda Cuevas, Fernando Luís Barroso da Silva

RNA and Transcriptomics

159

- Preliminary analysis of microRNAs and their pathway genes in the genome of *Globodera pallida*.
Carlos Bruno de Araujo, Júlia Silveira Queiroz, Caio Borges Melo, Matheus de Souza Gomes, Laurence Rodrigues do Amaral

- 160** Human riboswitch: are we close to predicting it?

Deborah Antunes, Fabio Passetti, Ernesto Raul Caffarena

- 161** miRNAs Expression and in silico Prediction of Targets Related with Resistant Exercise and Carbohydrate/Protein Supplementation

Adrielle Vieira de Souza, Miguel Mauricio Diaz, Olga Lucia Bocanegra, Renata Roland Teixeira, Maria Carolina Siqueira, Matheus de Souza Gomes, Fouad Salmen Espindola

- 162** Preliminary analysis of Dicer-like genes in Cucumber genome

Júlia Silveira Queiroz, Tamires Caixeta Alves, Núbia Carolina Pereira Silva, Matheus de Souza Gomes, Laurence Rodrigues do Amaral

- 163** Archaeal RNA polymerase pausing modeling and its gene expression control impacts

Danillo Cunha de Almeida e Silva, Tie Koide, Ricardo Zorzetto Nicoliello Vêncio

- 164** Transcriptome meta-analysis reveals the human organs evolution

Katia de Paiva Lopes, Ricardo Assunção Vialle, José Miguel Ortega

- 165** Transcriptome analysis of mice hearts infected with two strains of *Trypanosoma cruzi*: Insights into the parasite effects on the host gene expression

Tiago Bruno Rezende de Castro, Maria Cecília Campos Canesso, Mariana Boroni, Nayara Toledo, Carlos Renato, Égler Chiari, Andréa Mara Macedo, Glória Regina Franco

- 166** Comparative analysis of gene expression data between colorectal cancer cell lines with wild-type and silenced MMR genes

Cristóvão Antunes de Lanna, Nicole Scherer, Mariana Boroni

- 167** Functional and structural characterization of RBP42 in *Trypanosoma cruzi*

Daniela de Laet Souza, Daniela Ferreira Chame, Eddie Luidy Imada, Helaine Graziele Santos Vieira, Andréa Mara Macedo, Carlos Renato, Glória Regina Franco

- 168 RNA-binding proteins ALBA3 and DRBD3 characterization on *Trypanosoma cruzi* under gamma irradiation stress
Daniela Ferreira Chame,Daniela de Laet Souza,Eddie Luidy Imada,Helaine Graziele Santos Vieira,Andréa Mara Macedo,Carlos Renato,Glória Regina Franco
- 169 Occurrence of differential alternative splicing in the transcriptome of mice hearts infected with two strains of *Trypanosoma cruzi*
Nayara Toledo,Tiago Bruno Rezende de Castro,Glória Regina Franco,Carlos Renato,Égler Chiari,Andréa Mara Macedo
- 170 Long Noncoding RNAs in Patients with Dengue: Insights into Gene Regulation
Matheus Carvalho Bürger,Lucas Cardozo,Thiago Dominguez Crespo Hirata,Helder Takashi Imoto Nakaya
- 171 Transcriptome analysis of *Corynebacterium pseudotuberculosis* in an iron deficient environment
Izabela Coimbra Ibraim,Flavia Figueira Aburjaile,Thiago Luiz de Paula Castro,Núbia Seiffert,anne cybelle pinto gomide,Vasco A de C Azevedo
- 172 Annotation and analysis of the dynamics of splice acceptor sites in *Trypanosoma cruzi* under gamma radiation stress
Andre Luiz Martins Reis,Mainá Bitar,Helaine Graziele Santos Vieira,Dominik Kaczorowski,Andréa Mara Macedo,Carlos Renato,Glória Regina Franco
- 173 De novo assembly of *Trypanosoma cruzi* strain CL Brener transcriptome
Eddie Luidy Imada,Mainá Bitar,Máira R Rodrigues,Daniela Ferreira Chame,Helaine Graziele Santos Vieira,Andre Luiz Martins Reis,Michele Araújo Pereira,William Santos Prado,Dominik Kaczorowski,Andréa Mara Macedo,Carlos Renato,Martin Alexander Smith,Glória Regina Franco
- 174 Evaluating RNA single bulges with a mesoscopic model
Erik de Oliveira Martins,Gerald Weber
- 175 Study on the variation of RNA Secondary Structure prediction as a function of Thermodynamic Parameters
Rodolfo Vieira Maximiano,Gerald Weber
- 176 Analysis and comparison of force field of RNA using molecular dynamic simulations.
Rodrigo Bentes Kato,Jadson Claudio Belchior
- 177 Transcriptome profiling in *Leishmania amazonensis* promastigotes associated with virulence attenuation
Gabriela Flavia Rodrigues Luiz,Mariana Costa Duarte,Daniel Menezes-Souza,Ricardo Fujiwara,Eduardo Antonio Ferraz Coelho,Daniella Bartholomeu
- 178 De novo transcriptome assembly and comparative expression profiling of midgut tissues of four non-model insects
Rajesh Kumar Gazara,Christiane Cardoso,Daniel Bellieny Rabelo,Clélia Ferreira,Thiago Motta Venancio,Walter R. terra
- 179 Differential expression in colorectal cancer progression
Aline Duarte Gomes,Paulo Thiago Santos,Nicole Scherer,Mariana Boroni
- 180 Comparative transcriptome profiling of virulent and non-virulent *Trypanosoma cruzi* underlines a role of surface proteins during infection
Gabriela Flavia Rodrigues Luiz,Ashton Trey Belew,Rondon Pessoa de Mendonça-Neto,Edson Oliveira,Bruna Mattioli Valente,Rafael B. Polidoro,Ricardo T. Gazzinelli,Daniella Bartholomeu,Barbara A. Burleigh,Najib M. El-Sayed,Santuza Maria Ribeiro Teixeira
- 181 Acute Myeloid Leukemia gene co-expression networks and differential expression analysis in blood and bone marrow samples
Kendi Nishino Miyamoto,Diego Bonatto
- 182 Large scale transcriptional analysis of an animal model of seizures
Samara Damasceno,Cristiane de Souza Rocha,Iscia Teresinha Lopes-Cendes,Ana Lúcia Brunialti Godard
- 183 Determining the stability of DNA/RNA hybrid duplexes
Vivianne Basílio Barbosa,Erik de Oliveira Martins,Gerald Weber

- 184 Comprehensive profiling and characterization of *Arachis stenosperma* (peanut) and *Meloidogyne arenaria* (plant-root nematode) small-RNAs identified during the course of the infection
Priscila Grynberg, Larrisa A. Guimarães, Marcos Mota do Carmo Costa, Roberto Coiti Togawa, Ana Cristina M. Brasileiro, Patricia Messenberg Guimarães
- 185 Non-coding RNAs putatively acting as ceRNAs in embryonic stem cells
Raquel Calloni, Diego Bonatto
- 186 Optimized RNA nearest-neighbor enthalpy and entropy parameters as function of salt concentration
Izabela Ferreira, Elizabeth Jolley, Brent Znosko, Gerald Weber
- 187 Sequence-independent metagenomic analysis of animal viromes based on molecular characteristics of small RNAs
Lucio Rezende Queiroz, Eric Roberto Guimarães Rocha Aguiar, Roenick Proveti Olmo, João T. Marques
- 188 LncRNAPlant-Finder: a tool for prediction of long non coding RNAs in plants
Tatianne da Costa Negri, Pedro Henrique Bugatti, Priscila Tiemi Maeda Saito, Douglas Silva Domingues, Alexandre R. Paschoal
- 189 Transcriptional memory contributes to tolerance to multiple drought exposures in coffee (*Coffea canephora*) plants
Fernanda Alves de Freitas Guedes,
- 190 Mirtrons: computational feature analysis and miRNA comparison
Tamires Priscila da Costa, Douglas Silva Domingues, Alexandre R. Paschoal
- 191 Effect of Cy3 and Cy5 dyes on the hydrogen bonds of oligonucleotides
Pâmella Miranda de Moura, Luciana M. Oliveira, Gerald Weber
- 192 SNP discovery in the *Klebsiella pneumoniae* transcriptome after polymyxin B induction in combination with abiotic stresses using RNA-Seq technology
Thiago Cardoso Pereira Carneiro, Guilherme Loss de Moraes, Gisele Lucchetti, Guadalupe del Rosario Quispe Saji
- 193 Integrative analysis of transcriptomics and metabolomics data: adaptation of *Propionibacterium freudenreichii* to long-term survival in nutritional shortage
Flavia Figueira Aburjaile, Anderson Miyoshi, Artur Silva, Vasco A de C Azevedo, Yves Le Loir, Hélène Falentin
- 194 Single nucleotide variation analysis in microRNA target regions in colorectal cancer
Jéssica Noronha Blanco, Natasha Jorge, Fabio Passetti
- 195 Transcriptome analysis of high-temperature stress in yeast during industrial scale bioethanol production
Luciana Souto Mofatto, Osmar Vaz de Carvalho-Netto, Gleidson Silva Teixeira, Silvio Roberto Andrietta, Maria da Graça Stupiello Andrietta, Gonçalo Amarante Guimarães Pereira, Marcelo Falsarella Carazzolle
- 196 The assessment of the impact of small deletions within human protein domains using transcriptome data: a preliminary analysis.
Fernanda Cristina Medeiros de Oliveira, Gabriel Wajnberg, Fabio Passetti
- 197 Non-coding RNAs in the genus *Aeromonas*
Jean Carlos Machado da Costa, Maria Berenice R. Steffens, Alexandre Rossi Paschoal, Cynthia Maria Teles Fadel Picheth, Fabio de Oliveira Pedrosa
- 198 Altered gene expression by control unspecific dsRNAs: an inquiry
Sandra Grossi Gava, Naiara Cristina Clemente Santos Tavares de Paula, Anna Christina de Matos Salim, Flávio Marcos Gomes Araújo, Guilherme Oliveira, Marina de Moraes Mourão
- 199 PIWI-interacting RNA and small nucleolar RNA signatures of smokers and non-smokers in lung adenocarcinoma
Natasha Jorge, Gabriel Wajnberg, Benílton Carvalho, Carlos Gil Ferreira, Fabio Passetti
- 200 De novo transcriptome assembly of the extremophile plant *Calotropis procera*
Ana Maria Benko-Iseppon, Rebeca Rivas, João Pacifico Bezerra Neto, Valesca Pandolfi, Maria Reis Velois Coêlho, Mauro Guida dos Santos
- 201 TFBS prediction in sugarcane using binding sites prediction from PlantTFDB server
Mauro de Medeiros Oliveira, Danielle Izilda Rodrigues da Silva, Alan Durham, Glauzia Souza Mendes

- 202 Integrative bioinformatics data analysis of Nile Tilapia microRNAs
Luiz Augusto Bovolenta,Danillo Pinhal,Simon Moxon,Arthur Casulli de Oliveira,Pedro Gabriel Nachtigall,Marcio Luis Acencio,Cesar Martins,Ney Lemke
- 203 Tityus serrulatus venom gland: new sodium channel toxins through RNA-Seq
Ana Paula Vimieiro Martins,Flávia de Faria Siqueira,Evanguedes Kalapothakis
- 204 PiRNA signatures of adjacent to tumor tissue as potential biomarkers of gastric carcinogenesis
André Mauricio Ribeiro dos Santos,Tatiana Vinasco Sandoval,pablo pinto,Amanda Vidal,Arthur Ribeiro-dos-Santos,Paulo Assumpção,Mônica Assumpção,Sâmia Demachki,Sidney Santos,ÂNDREA KELY CAMPOS RIBEIRO DOS SANTOS,Sandro Jose de Souza,Fabiano Moreira
- 205 Analysis of the lincRNA transcriptome in the accessory olfactory system
Antônio Pedro de Castello Branco da Rocha Camargo,Marcelo Falsarella Carazzolle,Fabio Papes
- 206 Identifying gene clusters in the genome of Trypanosoma cruzi
Willian Santos Prado,Andre Luiz Martins Reis,Tiago Bruno Rezende de Castro,Glória Regina Franco
- 207 Molecular diversity of the venom gland from Peruvian scorpion Hadruroides lunatus revealed by transcriptome analysis.
Thiago Mafra Batista,
- 208 Trypanosoma cruzi coding transcriptome in response to gamma radiation
Michele Araújo Pereira,Eddie Luidy Imada,Priscila Grynberg,Helaine Graziele Santos Vieira,Dominik Kaczorowski,Andréa Mara Macedo,Carlos Renato,Glória Regina Franco
- 209 Combined genome guided and long reads assembly of the Coffea arabica transcriptome
Pâmela Marinho Rezende,Thales Henrique Cherubino Ribeiro,Fernandes-Brum C. N.,Schumacher P. V.,Ferrara-Barbosa B. C.,Chalfun-Junior A.
- 210 Assembly, identification and characterisation of sugarcane transcripts
Pâmela Marinho Rezende,Thales Henrique Cherubino Ribeiro,Schumacher P. V.,Lima A. A.,Chalfun-Junior A.
- 211 Caprin-1 binding profile to target RNAs via enhanced CLIPseq
Felipe Ciamponi,Natacha Migita,Eric Van Nostrand,Michael Lovci,Stefan Aigner,Laura Alonso,Gene Yeo,Katlin Massirer
- 212 Comparison of the Expression profile between embryogenic and non-embryogenic Coffea arabica L. calli through RNA-Seq data analyses using combination of DE algorithms
Thales Henrique Cherubino Ribeiro,Wesley Pires Flausino Máximo,Kalynka Gabriella do Livramento,Anderson Tadeu Silva,Chalfun-Junior A.,Luciano Vilela Paiva
- 213 Genome-wide identification and in silico characterization of microRNAs and their targets in Ananas comosus L.
Thales Henrique Cherubino Ribeiro,Pâmela Marinho Rezende,Laurence Rodrigues do Amaral,Matheus de Souza Gomes,Chalfun-Junior A.
- 214 Lsm-bound antisense RNAs play role in Halobacterium salinarum NRC-1 transposition regulation
Alan Péricles Rodrigues Lorenzetti,José Vicente Gomes-Filho,Lívia S. Zaramela,Felipe ten Caten,Ricardo Zorzetto Nocoliello Vêncio,Tie Koide
- 215 Non coding RNAs in Coffea canephora genome: Identification by similarity.
Samara Mireza Correia de Lemos,Alexandre Rossi Paschoal,Douglas Silva Domingues
- 216 Distinguishing coding and non-coding RNA sequences and improving its functional annotation using machine learning approaches
Thaís de Almeida Ratis Ramos,Daniel Miranda de Brito,Raúl Arias-Carrasco,Leonardo Vidal Batista,Thaís Gaudencio,Vinicius Maracaja Coutinho
- 217 Proposal of a data mining pipeline to improve bacterial small RNA prediction
Fabio Ivan Reinoso Vilca,Sabrina de Azevedo Silveira,Fabio Ribeiro Cerqueira
- 218 Differential Gene Expression Analysis of Placentas from *Mus musculus* Exposed to Different Stress Conditions
André Rocha Barbosa,Ana Carolina Tahira,Helena Brentani

- 219 Transcriptional landscape of *Paracoccidioides brasiliensis*: an isolate presenting no dimorphism shift
Luciana M. Oliveira, Christooher Desjardins, Jerônimo C. Ruiz, Viviane Alves, Ludmila M. Baltazar, Patrícia C. Santos, Christina Cuomo, Patrícia S. Cisalpino

Software Development and Databases

220

- 220 R script to HLA epitope predictor based in matrix frequency: training and performance comparisons
Alessandra Lima da Silva, Leandro Martins de Freitas
- 221 Patent Mining as a Tool for Innovation Planning and Biodiversity Access: Technologies of Açaí Fruit (Euterpe oleraceae Mart).
Foued Salmen Espindola, Heitor Cappato, Letícia de Castro Guimarães, Fabiana Regina Grandeaux de Melo
- 222 mirhunt: an approach to predict microRNA binding sites using different prediction tools
Jaqueline Ramalho, Michelle Almeida da Paz, Iane de Oliveira Pires Porto, Celso Teixeira Mendes-Junior, Erick da Cruz Castelli
- 223 Alien: A tool for handling sequence alignments
Dhiego Souto Andrade,
- 224 Development a Predictor of Aggregation-protein using Supporting Vector Machine
Carlos Alves Moreira, Prof. Dr. Luis Paulo Barbour Scott
- 225 Towards Transparent and Reproducible Bioinformatics Analyses: the EPIGEN-Brazil Scientific Workflow
Gilderlanio Santana de Araújo, Wagner Magalhaes, Eduardo Martin Tarazona Santos, Maíra R Rodrigues
- 226 LabControl: a LIMS software to manage microbial data
Mariana Teixeira Dornelles Parise, Joarley Ferreira dos Santos, Aristóteles Góes-Neto, Daniela Arruda Costa, Anne cybelle pinto gomide, Gabriel da Rocha Fernandes, Vasco A de C Azevedo
- 227 Noninvasive prenatal paternity determination by SNPs and microhaplotypes
Jaqueline Yu Ting Wang,
- 228 Novel bioinformatic approaches for viral discovery from NGS data
Liliane Santana Oliveira Kashiwabara, João M. P. Alves, Dolores U. Mehnert, Alan Durham, Paolo M. A. Zanotto, Arthur Gruber
- 229 Software Assessment for Prediction of Gene Clusters: An Analysis in silico with Cyanobacteria of Chroococcales Order
Danielle Costa Carrara Couto, Vanessa C. Rezende, Alex R. J. Lima, Felipe Carrara Couto, Leonardo Teixeira Dall'Agnol, Evonnildo C. Gonçalves
- 230 Brimer: A Web System for Managing Primers
Danielle Costa Carrara Couto, Aryane P. Vilhena, Felipe Carrara Couto, Leonardo Teixeira Dall'Agnol, Hivana P. M. B. Dall'Agnol, Evonnildo C. Gonçalves
- 231 Network Algorithm To Relatedness Analysis (NAToRA)
Thiago Peixoto Leal, Mateus Gouveia, Gilderlanio Santana de Araújo, Maíra R Rodrigues, Marilia Sciliar, Eduardo Martin Tarazona Santos
- 232 HaploCYP: a software for CYP2D6 genotyping and phenotype prediction
Michele Araújo Pereira, Raony Guimaraes Corrêa Do Carmo Lisboa Cardenas, Elvis Cristian Cueva Mateo, Alessandro Clayton de Souza Ferreira, Maíra Cristina Menezes Freire
- 233 Text mining for HPC
Bruna Pierreck Moura, Adriano Barbosa-Silva, Reinhard Schneider, Sarah Diehl, Ana Christina Brasileiro-Vidal, Ana Maria Benko-Iseppon
- 234 Database Model for Fish Collection
Maria Fernanda Hussni, Nicolás Valentín Molina Terra, Marcelo Cesar Pinto, Luiz Henrique Garcia Pereira
- 235 A comprehensive database of mirtrons knowledge
Bruno Henrique Ribeiro da Fonseca, Douglas Silva Domingues, Alexandre R. Paschoal

- 236 Linking microbial community composition and potential functional roles using shotgun metagenomic libraries
Laura Rabelo Leite, Julliane Dutra Medeiros, Francislon Silva de Oliveira, Victor Satler Pylro, Sara Cuadros Orelana, Guilherme Oliveira, Gabriel da Rocha Fernandes
- 237 DevOps cloud-computing environment to perform virtual screening
Leo Rodrigues Biscassi, Rodrigo Antônio Faccioli, Paulo Eduardo Ambrósio
- 238 BDGF: a database and web-based information retrieval system for genotype and phenotype
Fábio Danilo Vieira, Danilo Gomes de Moura, Diego Félix da Silva, Roberto Hiroshi Higa, Adhemar Zerlotini Neto
- 239 PFStats: A tool for protein analysis by decomposition of residue coevolution networks and amino acid reduced alphabets applications
Nélio José da Fonseca Júnior, Lucas Bleicher, Marcelo Querino Lima Afonso
- 240 Classifiers for patients with breast cancer according to the neoadjuvant chemotherapy sensitivity
Pedro Kássio Ribeiro Matos Loureiro de Carvalho, Thiago de Souza Rodrigues
- 241 Impact of genomic RNA structure and non-coding RNAs in Zika virus neuropathogenesis
Fernanda Luz Castro, Raúl Arias-Carrasco, Yessenia Vásquez-Morán, Artur Lopo de Queiroz, Helder Takashi Imoto Nakaya, Renato Santana de Aguiar, Vinicius Maracaja Coutinho
- 242 DNAShot: an application to Blast DNA Sequence from photos using smart-phone
RICARDO VOYCEIK,
- 243 Evaluation of predictor programs of genomic islands
Diônata Willian Augusto, Antonio Camilo da Silva Filho, Izabella Castilhos Ribeiro dos Santos Weiss, Paulo Afonso Bracarense Costa, Jeroniza Nunes Marchaukoski
- 244 miRNAPath II: platform to identify miRNAs targets and pathways regulated by miRNAs
Natalia Baptista Cruz, Jessica Rodrigues Plaça, Wilson Araújo da Silva Jr
- 245 A graphical tool for data integration and analysis of complex diseases
Sérgio Nery Simões, João Carlos Pandolfi Santana, Ana Rubia Ramos Vicente, David Correa Martins-Jr
- 246 Identifying Alternative Splicing Events in RNAseq data using De Bruijn Graphs and Bloom Filters
Ricardo Medeiros da Costa Junior, André Yoshiaki Kashiwabara
- 247 Capturing experimental detail in a paperless environment – Scarab, an Electronic Lab Notebook developed and used at the Structural Genomics Consortium
Lucas Ferreira, Katlin Massirer, Opher Gileadi, Brian Marsden
- 248 Have you ever wanted to learn about the cladistics origin of operons? Take our TAXI (Taxonomic Innovations)
Lucas Ferreira, José Miguel Ortega
- 249 NCBI NR Protein Database Clustered by Homology Inference
Aryel Marlus Repula de Oliveira, Roberto Tadeu Rattz
- 250 Meta-analysis of Japanese Toxicogenomics data: differences between in vivo and in vitro models
Carlos Biagi Jr, Richa Batra, Jan Baumbach, Jose Luiz Rybarczyk Filho
- 251 A database for comparative analysis of paradigms for prospecting contacts in protein-protein interfaces
Pedro Magalhães Martins, Vinícius Diniz Mayrink, Sabrina de A. Silveira, Carlos Henrique da Silveira, Leonardo Henrique França de Lima, Raquel Melo Minardi
- 252 The Polyploid Gene Assembler (PGA)
Leandro Costa do Nascimento, Gonçalo Amarante Guimarães Pereira, Marcelo Falsarella Carazzolle
- 253 An approach for constructing a database of manually curated contacts in proteins
Marcos F M Silva, Pedro M Martins, Diego César Batista Mariano, Isabela Pastorini, Naiara Pantuza, Raquel Melo Minardi
- 254 Using Data Marts to Select Related Research Articles: A Case Study for the Prioritization of Drug Targets
Marlon Amaro Coelho Teixeira, Maria Claudia Cavalcanti, Floriano Paes Silva Junior, Kele Teixeira Belloze
- 255 ATENA: A decision support system for classifying genetic variants and clinical diagnosis
Renata Correia de Andrade, Marcel Caraciolo, Joao Bosco Oliveira, George de Vasconcelos Carvalho Neto

- 256 The importance of an adequate soft-clip based approach on bioinformatics pipelines for multiplex targeted next-generation sequencing
George de Vasconcelos Carvalho Neto, Renata Correia de Andrade, Marcel Caraciolo, Joao Bosco Oliveira, Rodrigo Bertollo
- 257 visGReMLIN: An interactive strategy to visualize common substructures in protein-ligand interaction
Vagner Soares Ribeiro, Charles A. Santana, Fabio Ribeiro Cerqueira, Alexandre Victor Fassio, Carlos H. da Silveira, Raquel Melo Minardi, Sabrina de Azevedo Silveira
- 258 An automated method for the identification of Dengue, Zika, Yellow Fever and Chikungunya virus species and genotypes
Luiz Carlos Junior Alcantara, Nuno Rodrigues Faria, Marta Giovanetti, Vagner Fonseca, Maria Inés Restovic, Murilo Freire, Túlio de Oliveira
- 259 Dugong: a Docker image, inspired on Ubuntu Linux, designed to enhance reproducibility and replicability during computational analyses of biological data
Fabiano Menegidio, Luiz Nunes

Systems Biology and Networks

260

- 260 CEMiTool: Co-expression Modules Identification Tool
Pedro de Sá Tavares Russo, Gustavo Rodrigues Ferreira, César Augusto Prada Medina, Matheus Carvalho Bürger, Lucas Cardozo, Luciane Schons Fonseca, Thiago Dominguez Crespo Hirata, Gonzalo Sepúlveda Hermosilla, Vinicius Maracaja Coutinho, Helder Takashi Imoto Nakaya
- 261 Cancer immunology of Cutaneous Melanoma: A Systems Biology Approach
Mindy Muñoz Miranda, Pedro de Sá Tavares Russo, Gustavo Rodrigues Ferreira, Helder Takashi Imoto Nakaya
- 262 Global coexpression analysis of human protein-coding genes
Katia de Paiva Lopes, Francisco José Campos-Laborie, Ricardo Assunção Vialle, José Miguel Ortega, Javier De Las Rivas
- 263 Meta-dimensional analysis in gene network inference and gene prioritization associated to complex diseases
Carlos Eduardo Marchi, David Corrêa Martins Junior, Fábio Marchi
- 264 Logical Modeling of Cellular Senescence Induced by DNA Damage and TGFbeta signaling
José Carlos Merino Mombach, Veronica Venturini Rossato
- 265 Saccharomyces cerevisiae Protein-Protein Interaction Network Reconstruction to Study Ethanol Tolerance
Ivan Rodrigo Wolf, Lauana Fogaça, Leonardo Nazário de Moraes, Rejane M. T. Grotto, Rafael P. Simões, Guilherme Valente
- 266 A logical model for the bimodal p53 switch in cell-fate control
José Carlos Merino Mombach, Maria V. C. Issler
- 267 Pattern Recognition in genomic sequences: A case of study using complex networks
Isaque Katahira, Fabricio Martins Lopes
- 268 Influence of a high-fat diet in the cerebellar tissue of Cockayne Syndrome mice
Gabriel Baldissera, Kendi Nishino Miyamoto, Diego Bonatto
- 269 ANOVA-like method for differential correlation of multiple networks analysis of biological data
Vinicius Jardim Carvalho, Suzana Siqueira Santos, Amanda Pereira de Souza, Adriana Grandis, Andre Fujita, Marcos Silveira Buckeridge
- 270 Using Systems Biology to understand Immunosenescence
Fernando Marcon Passos, Helder Takashi Imoto Nakaya, Pedro de Sá Tavares Russo, Matheus Carvalho Bürger, Thiago Dominguez Crespo Hirata
- 271 Construction of metabolic map in lead poisoning
Iara Dantas de Souza,
- 272 RTNsurvival: An R package for survival analysis from regulatory networks
Vinicius Chagas, Clarice Groeneveld, Kelin G. Oliveira, Mauro Antônio Alves Castro

- 273** Interactome analysis of FGFR2 – a potential therapeutic target in breast cancer.
Kelin Gonçalves de Oliveira,Mauro Antônio Alves Castro
- 274** Understanding transcriptional strategy for Inositol pathway in soybean root dehydration stress tolerance
João Pacifico Bezerra Neto,José Ribamar Costa Ferreira-Neto,Ederson Akio Kido,Manassés Daniel da Silva,Ana Maria Benko-Iseppon,Mauro Guida dos Santos
- 275** Gene Regulatory Network Modeling for Mycelium-to-Yeast Transition of Paracoccidioides brasiliensis
Luciane Sussuchi da Silva,Célia Maria de Almeida Soares,Alexandre Melo Bailão,Clayton Luis Borges,Juliana Alves Parente Rocha,Juliano Paccez,Roosevelt Alves da Silva,Gustavo Goldman,Luis Anibal Diambra,Maristela Pereira
- 276** Integrating omics data from xylose-fermenting yeast using network dynamic modeling for bioethanol production
Lucas Miguel de Carvalho,Gabriela Vaz de Meirelles,Renan Augusto Siqueira Pirolla,Leandro Vieira dos Santos,Gonçalo Amarante Guimarães Pereira,Marcelo Falsarella Carazzolle
- 277** Design and Engineering of Synthetic Biological Systems for Medical Diagnosis
Francisco Schneider,Alexis Courbet,Christophe Nguyen,Marina Cardia Jardim,Liyan He,Laurence Molina,Liza F. Vilela,Patrick Amar,Franck Molina
- 278** An intuitive network-based approach to investigate clinical features among breast cancer subtypes
Andre Fonseca,Sandro José de Souza
- 279** SigNetSim : A web tool for modeling and analyzing quantitative biochemical networks
Vincent Noel,Marcelo S. Reis,Matheus H.S. Dias,Lulu Wu,Amanda S. Guimarães,Daniel F. Reverbel,Junior Barrera,Hugo A. Armelin
- 280** Hierarchical Model of the Ras-MAPK signalling pathway in mouse Y1 adrenocortical tumor cells
Vincent Noel,Marcelo S. Reis,Matheus H.S. Dias,Cecilia S. Fonseca,Layra L. Albuquerque,Fabio Nakano,Junior Barrera,Hugo A. Armelin
- 281** Analysis and Mining Onco-targets Breast through Ontology
Edgar Lacerda de Aguiar,Lissur Azevedo Orsine,Carlos Alberto Xavier Gonçalves,Marcos Augusto dos Santos,José Miguel Ortega
- 282** Identifying metabolic processes shared among genome-wide association studies for reproductive phenotypes in bovine
Pablo Augusto de Souza Fonseca,Luíza de Almeida Ferreira Diniz,Fernanda Caroline dos Santos,Izinara Rosse da Cruz,Maria Raquel Santos Carvalho
- 283** Gene network analysis of melanoma cancer development
Fabiano Sviatopolk Mirsky Pais,Diego Vinícius de Castro Pereira
- 284** Ancestrality and evolution of genes related with apoptosis
Rayson Carvalho Barbosa,Carlos Alberto Xavier Gonçalves,Lab Biodados
- 285** GEN3VA: Aggregation and Analysis of Gene Expression Signatures from Related Studies
Caroline D Monteiro,
- 286** Metabolic pathways involved in bovine temperament
Fernanda Caroline dos Santos,Pablo Augusto de Souza Fonseca,Izinara Rosse da Cruz,Luíza de Almeida Ferreira Diniz,Maria Raquel Santos Carvalho

1 | Organizing Committee

AB3C President: Glória R Franco (UFMG)

AB3C Vice President: Alan M Durham (USP)

BSB Chair: Sérgio Campos (UFMG)

AB3C Secretaries :

- Marcelo Brandão (Unicamp)
- Ney Lemke (Unesp)

AB3C Financial Department :

- Priscila Grynberg (Embrapa)
- Fábio Passetti (Fiocruz)

Poster Session Organizers :

- Mainá Bitar (UFMG)
- Nicole Scherer (INCA)

Paper Submission Organizers :

- Sérgio Campos (UFMG)
- Marcelo Brandão (Unicamp)
- Ney Lemke (Unesp)
- André Fujita (USP)
- Ronnie Alves (Université Montpellier, França)

Local Committee :

- Alan Durham (USP)
- André Fujita (USP)
- Arthur Gruber (USP)
- Ronaldo Fumio Hashimoto (USP)

2 | Introduction

The Brazilian Association of Bioinformatics and Computational Biology (AB3C) is a scientific society funded in July 12th 2004. Since its creation, AB3C has been responsible for the annual conference entitled “X-Meeting” which is the main Bioinformatics and Computation Biology event in Brazil. This year its 11th edition will be held in São Paulo, the biggest city in South America.

Bioinformatics is now a strategic area for Brazil and all Latin America and, therefore, it is also strategic to the development of Science, Technology and Economy. The X-Meeting is a Brazilian event with international reach which has an average of 400 participants. The Conference is an opportunity for students, researchers and companies to interact and difuse knowledge. The AB3C has been a pioneer society in the field of Bioinformatics in Brazil and we have a history of ten past very productive meetings.

3 | Abstracts

Design of chimeric antigens of Porcine Reproductive and Respiratory Syndrome Virus (PRRSV) through bioinformatics approaches: a rational model for the development of a diagnostic test

Jerusa Botelho Souza, Giuliana Loreto Saraiva, Jackson de Andrade Teixeira, Pedro Marcus Pereira Vidigal, Márcia Rogéria de Almeida

Laboratório de Infectologia Molecular Animal, Departamento de Bioquímica e Biologia Molecular, Núcleo de Análise de Biomoléculas, Universidade Federal de Viçosa, Viçosa, Minas Gerais

Brazil is one of the largest producers and exporters of swine meat in the world and therefore it is necessary a more stringent control of diseases affecting the swine herd. In this context, the Porcine Reproductive and Respiratory Syndrome Virus (PRRSV) is the important etiologic agent that causes significant economic losses. There is not official report of PRRSV in Brazil, but there is the possibility of its introduction by the marketing of matrices and semen from endemic countries and by contact with infected animals of border countries. Considering that the introduction of PRRSV in Brazil would have a major impact on the swine industry and on animal health, this research aims to produce a diagnostic test for PRRSV using chimeric antigens constructed with important regions for antibody recognition of both viral types (I and II). For this study, we constructed two databases with complete nucleotide sequences of GP5 and N genes of types 1 and 2 of PRRSV available in GenBank. The databases contemplated 1,768 complete sequences of GP5 and 91 complete sequences of N. The nucleotide sequences were aligned with MUSCLE algorithm and the alignments were reviewed, edited manually and served as the base line for phylogenetic and polymorphism analysis. Phylogenetic hypotheses were calculated by Bayesian Inference using Mr Bayes software version v3.1.2 and Bayesian Markov Chain Monte Carlo method (MCMC) in four runs with 100,000,000 generations for GP5 gene and 1,000,000 generations for the N gene. After viewing the phylogenetic trees, the nucleotide sequences were separated according to the groups formed (viral types 1 and 2) and translated for detailed investigation of polymorphisms using Python scripts. In this analysis, matrices containing the mutations identified in the datasets and their respective frequencies were calculated. With this information and from the alignment of amino acid sequences of each gene were obtained representative consensus sequences of each gene of each viral type using CLC Sequence Viewer version 7.6. After these analyzes, were performed predictions of transmembrane regions of GP5 protein to determine which fragments would be used in the construction of this synthetic gene. Each protein was built separately by joining the fragments of the two viral types with spacers. The genes of GP5 and N proteins were chemically synthesized and cloned into vectors for bacterial expression system. This project is standardizing the steps of expression and purification of recombinant proteins.

Acknowledgements: CAPES, CNPq, FAPEMIG e FUNARBE.

Prediction of microRNAs and miRNA pathway genes in *Solanum lycopersicum* and *Solanum pennellii*

Thaís Cunha de Sousa Cardoso¹, Tamires Caixeta Alves¹, Carolina Milagres Caneschi,
Douglas dos Reis Gomes Santana¹, Laurence Rodrigues do Amaral¹, Luiz Antônio
Augusto Gomes², Wilson Roberto Maluf², Matheus de Souza Gomes¹

¹Laboratory of Bioinformatics and Molecular Analysis – INGEB / FACOM, Federal University of Uberlândia, Campus Patos de Minas, Brazil, ²Department of Agriculture, Federal University of Lavras, Lavras, MG, Brazil

The cultivated tomato, *Solanum lycopersicum*, is one of the most important vegetable crops in global food and, next to the wild tomato *Solanum pennellii* are species widely used in developing cultivars. The study of the plant genomes has become a powerful tool to assist in the elucidation of the biological processes at the cellular level. One of the most important classes of small RNAs is microRNAs (miRNAs), acting on mRNA regulation in cells, inhibiting their translation and/or promoting its degradation. Computational methods have been applied extensively to identify novel miRNAs in different organisms. There are several proteins involved in the generation of miRNAs in plants, highlighting ARGONAUTE and DICER proteins which have key roles in the processing machinery of miRNAs. This study aimed to identify and characterize the genes involved in miRNA processing pathway as well as the miRNA molecules, their precursors and their target genes in the genome of *S. lycopersicum* and *S. pennellii* and also in next-generation sequencing. For the identification of the genes involved in miRNA pathway we used BLAST tools and reference genes available in NCBI, SolGenomics and Phytozomev11. We also performed domain and active site conservation analysis using PFAM and CDD databases and genome annotation files. The next-generation sequencing was held at Ion Personal Genome Machine®, using the Ion 318 chip. The targets of miRNAs were identified using the psRNATarget. We identified 65 proteins in the genome of *S. lycopersicum* and 109 in *S. pennellii* involved in small RNAs processing. Out of these proteins, 23 (*S. lycopersicum*) and 33 (*S. pennellii*) participate in the processing miRNAs pathway. In addition, we identified 342 different mature miRNAs, 226 precursor miRNAs distributed in 87 families, including 192 mature miRNAs not previously identified, belonging to 38 new families in *S. lycopersicum*. In *S. pennellii*, we found 338 mature miRNAs, 234 precursor miRNAs contained in 85 families. From the next generation sequencing, we identified 69 and 65 mature miRNAs distributed in 29 families and 28 in *S. lycopersicum* and *S. pennellii*, respectively. Furthermore, we identified 1310 different miRNA target genes in *S. lycopersicum* and 2772 in *S. pennellii*, suggesting important roles in plant development, regulation of hormonal response, defense against pathogens and other critical processes of biology, reproduction, and marketing of these species of tomato. Thus, our results expand the study of miRNAs in plants by providing new opportunities to understand essential in regulating processes based on miRNAs in tomato.

Funding support: FAPEMIG, CNPq and CAPES

*In silico genomic analysis of the endophytic bacterium *Bacillus amyloliquefaciens* 629*

Sant'Anna, Breno M. M.; Queiroz, Artur L.; Roque, Milton R. A.

Universidade Federal da Bahia, CPqGM/Fiocruz-Bahia, Universidade Federal da Bahia

The endophytic bacteria *Bacillus amyloliquefaciens* 629 isolated from *Theobroma cacao* L. was used to sequencing and genome annotation. Sequencing was performed using NGS Ion Torrent PGM platform (Life Technologies) 318 chip, the genome was assembled using SPAdes Genome Assembler version 3.5.0 and ordered by CONTIGuator 2.3. To evaluate and close gaps were performed manual curation with sequence alignment and editing tools, and also automatic annotation by RAST version 2.0 server. In order to build the phylogenetic tree, conserved sequences of 16S rRNA gene and rpoD were aligned and whole genome was used to DNA-DNA hibridization by ggdc (genome-to-genome distance calculator) and ANI (average nucleotide identity). Complementary analysis was carried out by Tetra Correlation Search (TCS). Here we describe the genome of *Bacillus amyloliquefaciens* 629, with 16 contigs, containing 3,903,367 bp, composed of 4,013 predicted genes, including 3,912 protein-coding sequences, 82 tRNAs and 19 copies of the genes for 5S, 16S, and 23S rRNA. Our comparative genome analysis reveals a new classification to strain 629, in the plantarum subspecies group, of *Bacillus amyloliquefaciens* subsp. plantarum, that shows synonymia with *Bacillus velezensis*. We used a target approach including a plant-host interaction profile, involved in chemotaxis, adhesion, colonization and motility genes in comparative analysis, to understand the essential metabolic systems for endophytic bacteria. The strain 629 showed 132 singletons compared with other strains of *Bacillus velezensis*, 9 genomic islands were predicted and 440 shared genes with other genomes of endophytic bacteria. Some of these regions of the genome establish advantages in endophytism process, to aid in the adaptation and colonization of different environments and host plants.

Genotoxicity testing in-silico: quantification of the DNA mutation caused by the glycosidic bond hydrolysis

Bárbara Zanandreiz Siqueira Mattos, Gustavo Passini dos Santos, Anton Semenchenko

Centro Universitário Newton Paiva, Belo Horizonte MG - Brazil

The goal of the current work is to estimate the missense mutation rates caused by the hydrolysis of the glycosidic bonds between the bases and the DNA backbone. These mutations naturally occur inside the living cells but give origin to innumerable genetic problems, for example, different types of cancers. Computer simulation of the DNA replication and polymerization is used as a tool to measure the rates of this class of mutations. The simulation is built upon the hybrid model of the DNA replication fork that includes polymerases representation and Okazaki segments. Agent based model combined with the Markov chain representation of the polymerase operation represents the essential elements of the biochemical reactions of this process. The DNA polymerization is a stochastic process defined by the Markov chain, while 2nd order chemical kinetics of the glycosidic bonds hydrolysis is represented as stochastic interactions between the agents performing random walk (Brownian motion) in the two-dimensional representation of the chromosomal environment. The estimates of the mutation rates and the investigation of the genotoxicity of water demonstrate the application of this computational system. The influence of water concentration on the missense mutation rate is illustrated for a number of model parameters. The technique presented can be extended at almost no cost to a large class of small molecule genotoxicity testing. Also, the potential application of this tool and method to wider industrial usage is discussed within the context of the mandatory testing of the new chemical and pharmaceutical products for both human consumption and environmental impact.

Admixture Mapping of Brazilians Identifies New Obesity Susceptibility Loci

Hanaisa de Pla e Sant Anna¹, Marilia Scliar¹, Meddly L. Santolalla Robles¹, Thiago Peixoto Leal¹, Gilderlanio Araujo¹, Mateus Gouveia¹, Wagner Magalhaes¹, Fernanda Kehdy², Eduardo Martin Tarazona Santos¹

¹ Departamento de Biologia Geral, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, 31270-901, Belo Horizonte, Minas Gerais, Brazil; ² Fundação Oswaldo Cruz, 21041-210, Rio de Janeiro, Rio de Janeiro, Brazil

Human obesity reached epidemic proportions and it imposes a public health and economic challenge of our time. It has been estimated that >40% of the individual variability of BMI, a common measurement of obesity, is attributed to genetic components. However, the genetic basis of obesity remains mostly unexplained. Genome-Wide Association Studies have identified at least 97 loci associated with obesity that only explain ~2.7% of the BMI variability. In this context, an effective but overlooked source to identify new genetic variants associated with obesity is the admixed genomes of Brazilians (European, African and Native American parental populations) using an Admixture Mapping strategy. This approach considers the differences in the prevalence of obesity between parental populations to find genomic regions that have enhanced ancestry of the parental populations with greater prevalence of the trait. Using this approach, we are working on an Admixture Mapping of BMI in Brazilians. We analyzed data of >2.28 million SNPs of three Brazilian cohorts, being 1,309 children from Salvador, 3,736 adults from Pelotas and 1,442 elderly people in Bambui. We aim to expand the catalogue of genetic variants associated with obesity and gain a better understanding of obesity in several age groups. Besides, we also are generating a pipeline which will facilitate further studies of Admixture Mapping of other diseases in Brazilians and other admixed populations. We found a significant positive correlation of BMI with African ancestry ($\rho=0.032$, $p\text{-value}=0.048$) and a negative correlation with European ancestry ($\rho=-0.04$, $p\text{-value} = 0.034$) in Pelotas. In Bambui cohort, we observed the opposite results of Pelotas, a negative correlation of BMI with African ancestry ($\rho=-0.06$, $p\text{-value}=0.031$) and a positive correlation with European ancestry ($\rho=0.06$, $p\text{-value}=0.023$). In Salvador there were no significant correlations of BMI with African or European ancestry ($p\text{-value}=0.447$ and 0.456, respectively). Correlation of Native American ancestry with BMI was not observed in any of the three cohorts. We identified new obesity susceptibility loci on chromosomes 10, 13, 16 and 20. We found that loci at 10q22.1-3 and 13q12.3 are associated with obesity in the children cohort of Salvador, while loci at 16q12.1 are associated with the adult obesity in Pelotas and at 20p12.1-2 with the female adult obesity in Pelotas. Our study highlights the potential of Brazilian genome to gain a better understanding of the genetic basis of diseases and it also provides a pipeline that supports future Admixture Mapping studies in trihybrid populations.

Diagnostic metagenomics: a case study based on suspected dengue infection

LC Conteville, MA Marín, AMB de Filippis, RMR Nogueira, MCL Mendonça, ACP Vicente

Oswaldo Cruz Institute, Oswaldo Cruz Foundation

Dengue virus infects an estimated 50–100 million people annually worldwide and hyperendemic in Brazil, where the mortality associated to this infection has reached 12% in 2010. During epidemics in Brazil, 50% of the suspected cases remain not confirmed even by specific RT-PCR, NS1 antigen test, IgM and viral cell culture and some of them evolved to fatal cases. In this study, our aim was to analyze a set of fatal cases, in order to reveal, minimizing the bias of specific diagnosis, any infectious agent present in these cases. We applied metagenomic approaches using high-throughput sequencing data from an Illumina HiSeq 2500 run. We performed quality control of the reads with cutadapt and prinseq. Then, clean reads that mapped to the human genome (Hg38) using Bowtie 2 with default parameters were filtered. Taxonomic analysis were performed with Kraken, GOTCHA, SURPI, Metaphlan2, Taxoner and Blastn that identified three viruses in distinct samples that were subsequently confirmed by specific PCR: Parvovirus B19, Hepatitis G virus and Torque-teno virus. Considering the Parvovirus B19, it was possible to recover its complete genome (5.6 kb) and determine that it belongs to genotype 1A, which is the predominant genotype worldwide. Two pathogenic bacteria were identified in four other samples: *N. meningitidis* serogroup C (n=2) and *Streptococcus pneumoniae* (n=2). Both bacteria have been causing outbreaks and epidemics in Brazil and infections with the former having high mortality rate. Therefore, metagenomics and bioinformatic analysis are an useful approach to reveal unpredictable epidemiological scenarios and should be applied in global emerging pathogen surveillance programs.

Detection of Functional Analogous Enzymes in the Human Metabolism

Piergiorgio RM, Guimarães ACR, Catanho M

*Fiocruz, Instituto Oswaldo Cruz, Laboratório de Genômica Funcional e Bioinformática,
Av. Brasil 4365, Manguinhos, Rio de Janeiro, 21040-900, RJ, Brazil*

Since enzymes catalyze almost all chemical reactions that occur in living organisms, it is important that genes encoding such activities are properly identified and functionally characterized. Several studies suggest that the fraction of enzymatic activities in which multiple events of independent origin have taken place during evolution is substantial. However, this topic is still poorly explored, and a comprehensive investigation of the occurrence, distribution and implications of these events, involving organisms whose genomes have been completely sequenced, has not been done so far. Fundamental questions, such as how analogous enzymes originate, why so many events of independent origin have apparently occurred during evolution, and what are the reasons for the coexistence in the same organism of distinct enzymatic forms, remain unanswered. In this context, the purpose of this project is to investigate the biological importance and the evolutionary role of functional analogous enzymes identified in metabolic pathways annotated in the human genome. A computational pipeline developed by our group (AnEnPi) was used to predict putative analogous enzymes employing protein sequences available in public databases (KEGG). The predicted functional analogy instances were confirmed by mining in Pfam, SUPERFAMILY and PDB databases for domain, folding and 3D structure information concerning the enzymes implicated. Using KEGG and Reactome databases as references, the predicted analogous enzymes were mapped in human metabolism. Altogether, we were able to detect convergence in 31 enzymatic activities (represented by EC numbers) belonging to 51 distinct processes and metabolic pathways from KEGG's reference maps. The genomic coordinates of the genes encoding these predicted analogous enzymes showed that these genes are dispersed throughout the human genome, found in 21 chromosomes. We selected Biliverdin Reductase analogs for further analyses. We found that, despite being considered isoenzymes, the two Biliverdin Reductase forms encoded in the Human genome, BLVRA and BLVRB, share remarkably low sequence similarity, among several other relevant differences, such as: BLVRA can interact with the DNA and regulated the gene expression, while no regulatory function has been described for BLVRB form so far. BLVRA is a component of the insulin signaling pathway and it is possible that both enzymes act on different isomers of biliverdin. Our findings suggest that the coexistence of multiple enzymatic forms in the Human genome might not be interpreted as functional redundancy. Instead, these enzymatic forms seem to be implicated in distinct (and probably relevant) biological roles.

Acknowledgments We wish to thank Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Programa Estratégico de Apoio à Pesquisa em Saúde (PAPES), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ), and Plataforma de Bioinformática Fiocruz RPT04-A/RJ for their support.

tRNA array genomic survey unravels their presence and structure in Mycobacteria

SM Morgado, ACP Vicente, MFA Marín

Laboratório de Genética Molecular de Microrganismos, IOC, FIOCRUZ

The tRNA arrays are genomic regions with highly density of tRNA found in eukaryotes and prokaryotes. Considering the importance of tRNAs in the translation machinery, these units influence the organism lifestyle. A previous *in silico* analysis showed their abundance in Firmicutes and conversely, rare among Actinomycetes. In the context of *Mycobacterium* genomic project, we sequenced several *Mycobacteria* genomes and analyzed them together with genomes from this genus. One of our focus was tRNA arrays identification. Currently, there is no tools to apply in the tRNA arrays identification. Therefore, we developed an *in house* perl script to identify such structures. The tRNA array units were defined as genomic regions containing at least 20 tRNA genes with a minimal tRNA gene density of 2 tRNA genes per kb. We identified tRNA arrays in several Actinomycetes genomes. These tRNA arrays contain from 21 to 62 tRNAs and are present in several *Mycobacterium* species, including *M. tuberculosis*. They were most abundant and diverse in *M. abscessus* complex. Same arrays in terms of composition and synteny were found intra and inter *Mycobacterium* species. However in *M. abscessus*, we identify the occurrence of diverse tRNA arrays, considering composition and synteny. The analysis based on the current script allowed the identification of distinct tRNA arrays in *Mycobacterium* genomes, raising questions about the impact of these structures in the biology of these organisms. The implementation of this script in genome pipelines would improve annotation as well as comparative genomic analyses of such structures.

Supported by: IOC, FIOCRUZ, CAPES

The first complete genome sequence of *Streptococcus dysgalactiae* subsp. *dysgalactiae* an emerging fish pathogen

Alexandra A. Urrutia Zegarra, Felipe L. Pereira, Fernanda A. Dorella, Alex F. Carvalho, Gustavo Morais Barony, Carlos A. G. Leal, and Henrique C. P. Figueiredo

Federal University of Minas Gerais, Belo Horizonte, MG, Brazil

Streptococcus dysgalactiae subsp. *dysgalactiae* (SDD) is a Gram-positive cocci, it autoaggregates in saline, forms long chains in growth medium, it is catalase negative and α -hemolytic on blood agar. In 2002, it caused the first outbreak in southern Japanese farms. During the subsequent years fish farms in the country suffered huge losses. In Brazil, outbreaks of streptococcosis are common in the freshwater fish species Nile tilapia, *Oreochromis niloticus* (L.). In 2007, the first disease outbreak caused by SDD was spotted in Ceará state. The disease has spread worldwide and despite its increasing clinical and economic significance up until the moment, none SDD genome was fully sequenced. Therefore, considering the importance of a complete genome to characterize this fish pathogen strategy, a next-generation sequence genome initiative was managed. To obtain the SDD genome the sample was isolated from an overnight culture with the Maxwell 16 tissue DNA purification kit using the Maxwell 16 system (both from Promega, USA). A first run was conducted on the Ion Torrent PGM™ sequencing system (Life Technologies, USA) using a 200bp (~ 300-fold coverage) fragment library kit. However, as it resulted in an overly fragmented assembly, another runs were performed using a 400bp (~870-fold coverage) fragment library kit and a 400bp (~ 107 fold coverage) mate-pair kit with an insert of 6kbp. Additional runs were conducted on the Illumina® MiSEQ sequencing system using paired-end 2x150bp (~638-fold coverage) and mate-pair (~658-fold coverage), with an insert of 6kbp. Yet, as no improvements were reached in the assembly fragmentation matter an optical map was acquired. The sequences were assembled with SPAdes 3.8.0, and Newbler 2.9 software, the assembly with higher N50 was selected and aligned with the Optical Map (OpGen Inc, USA) in order to verify the orientation and start scaffolding. Additionally, CONTIGuator software and the assembly_graph text file from the assembly output were used for further scaffold construction. Initially 167 contigs were obtained with an N50 value of 26,993bp and the largest contig with a 141,256bp length size and a ~44% of whole genome map (WGM) coverage. The first scaffolds constructed were used as input in a new assembly, this strategy lead to a better N50 (28,066bp) and fewer contigs (148). The procedure was repeated and ~52% of WGM coverage was reached. Currently, 84% coverage of the WGM was reached and gap filling with CLC Genomics Workbench 7 (Qiagen, USA) still in process. The present study empowers the use of optical mapping as a tool in the assembly of highly repetitive genomes. Further results as the first SDD complete genome announcement are expected.

This study was supported by the MAPA; FAPEMIG; CNPq and the INCT.

Assembly, annotation and comparison of *Corynebacterium pseudotuberculosis* lineages

Doglas Parise¹, Thiago de Jesus Sousa¹, Mariana Teixeira Dornelles Parise¹, Adrian Valentín Muñoz Bucio², Felipe Luiz Pereira¹, Fernanda Alves Dorella¹, Efrén Díaz Aparicio², Henrique Figueiredo¹, Daniela Arruda Costa¹, Vasco Ariston de Carvalho Azevedo¹

*Federal University of Minas Gerais*¹, *National Autonomous University of Mexico*²

Corynebacterium pseudotuberculosis (*Cp*) is a pathogenic bacterium that belongs to CMNR group (*Corynebacterium*, *Mycobacterium*, *Nocardia* e *Rhodococcus*). This group presents high CG content (46 – 74%) and cell wall composed of peptidoglycan, arabinogalactan and mycolic acids. Such bacterium is the etiological agent of caseous lymphadenitis in small ruminants and can affect other mammals as horses, buffaloes, camels and even humans. This work aims to characterize six *Cp* linages of both biovars (*ovis* and *equi*), isolated from Mexico. A key feature concerning those strains is that it is the first time *Cp* biovar *equi* is isolated from Mexico and any *Cp* isolated from this country is sequenced. The lineages *Cp* MEX1 and *Cp* MEX9 were isolated from goats, *Cp* MEX25 and *Cp* MEX29 were isolated from sheep and *Cp* MEX30 and *Cp* MEX31 were isolated from horses. The sequencing was performed in AQUACEN (UFMG) laboratory in Ion Torrent platform with a 400 base pairs (bp) fragment library kit. It generated an amount of data varying from 210,064,890 to 316,695,111 bp with coverage varying from 88.75-fold to 135.48-fold and medium phred quality varying from either 23 or 27 to each sequencing. The assembly methodology was based on hybrid strategy, performing *ab initio* assembly and contigs alignment through reference. Newbler 2.9, Mira 3.9 and SPAdes 3.6.0 assemblers were used and the best assembly was chosen considering the following criteria: number of contigs, genome size, maximum and minimum contig size, and N50. The selected assemblies varied from six to 33 contigs and were aligned against a reference genome utilizing CONTIGuator software. To obtain complete genomes the remaining gaps were filled using CLC software. Structural and functional annotation was performed in the next step, through automatic annotation and manual curation. The first was performed using RAST pipeline and an *in house* script to transfer the annotation of a previously curated genome, after this process both outputs were merged and manually curated. Until this moment, all strains except *Cp* MEX1 have already been annotated and curated. Such genomes presented ~2.3 mega bases of size, GC content of ~52%, ~2000 CDs, 1-77 pseudogenes, 12 rRNAs and 48-49 tRNAs. The linages *Cp* MEX9 and *Cp* MEX25 are available in NCBI, with the respective accession numbers: NZ_CP014543.1 and NZ_CP013697.1. As perspectives of this work, it is intended to finish the annotation and manual curation of *Cp* MEX1; deposit all strains and perform a comparative genomics study with them.

Identification of Specific Enzymes in the Comparison between *Fusarium oxysporum* and *Arabidopsis thaliana*

Larissa Catharina Costa¹ (lcosta@cdts.fiocruz.br), Nicolas Carels¹

¹*Laboratório de Modelagem de Sistemas Biológicos, National Institute of Science and Technology for Innovation in Neglected Diseases (INCT/IDN), Centro de Desenvolvimento Tecnológico em Saúde (CDTS), Fundação Oswaldo Cruz (Fiocruz), Rio de Janeiro, Brasil.*

The genus *Fusarium* currently accounts for more than 300 species and is composed of filamentous fungi including many crop pathogens. Species from genus *Fusarium* are fungi widely distributed in soil, plants and in different organic substrates. These mycotoxin-producing fungi cause diseases responsible for significant global economic losses and can even be opportunistic pathogens for humans. The identification of specific enzymes essential to the fungus metabolism might improve the identification of mechanisms for the control of fusariosis. Thus, we performed a comparison of non-redundant enzymatic activities between *Fusarium oxysporum* (879) and *Arabidopsis thaliana* (1044) using the list of Enzyme Commission numbers (ECs) of the Kyoto Database Encyclopedia of Genes and Genomes (KEGG). We recovered the identifiers and protein sequences of the enzymes that are specific to *F. oxysporum*. As the result of this comparison, we found that 248 ECs were only present in the fungus. The list of 248 ECs corresponded to 339 access numbers and 319 non-redundant protein sequences of *F. oxysporum*. The identification of specific enzymes of *F. oxysporum* is important since these enzymes may be participating in key pathways for the fungus survival. The success for parasite control of inhibiting specific enzyme activities in *F. oxysporum* will depend on (i) whether they are involved in essential pathways; (ii) alternative routes exist for these pathways; (iii) the level of disorders of the target inhibition can bring to the fungus survival; and (iv) the consequences of target inhibition to the host plant. Of course, the ideal target candidates to be inhibited are those that will not entailed any deleterious effect to the host plant.

Funding Support: CAPES and INCT-IDN (CNPq).

A comparative *in silico* linear B-cell epitope prediction for South American and African *Trypanosoma vivax* strains

Rafael Lucas Muniz Guedes¹, Carla Monadelí Filgueira Rodrigues², Luiz Gonzaga Paula de Almeida¹, Paola Minoprio³, Marta M.G. Teixeira², Ana Tereza Ribeiro de Vasconcelos¹

¹Laboratório Nacional de Computação Científica (LNCC), Av. Getúlio Vargas, 333, Petrópolis, RJ, Brazil.

²Department of Parasitology, University of São Paulo, Av. Prof. Lineu Prestes 1374, São Paulo, SP, Brazil.

³Department of Infection and Epidemiology, Institut Pasteur, Paris, France

Single-celled parasitic protists from the Kinetoplastida order are the etiological agents of trypanosomiasis, an important neglected disease that affects humans, domestic and wild animals world-wide. Animal tripanosomiasis, also known as Nagana, are caused by different species, i.e. *Trypanosoma vivax*, *Trypanosoma congolense* and *Trypanosoma brucei brucei*, being the former the most prevalent in livestocks in West Africa. *T. vivax* is disseminated across Africa and South America, infecting diverse mammals like cattle, sheep and goats leading to a profound economical impact in agriculture. Linear B-cell epitopes are predictable specific peptides recognized by host's antibodies triggering immune response. Due to the difficulty of maintaining *T. vivax* under laboratory conditions, little is known about its immunogenicity. Here we have combined several bioinformatic tools in order to select the best *in silico* linear B-cell epitope candidates for improving serodiagnosis and serotyping. A representative dataset of *T. vivax* strains was prepared with transcriptomic data from bloodstream forms from the Western African (Til: *T. vivax* IL1392, from Kenya) and two South American (Tsp: *T. vivax* Lins, from São Paulo state, Brazil, sequenced in the present work and Tvv: *T. vivax* LIEM-176, from Venezuela) isolates. The annotated genome of another Western African strain (Tvi: *T. vivax* Y486, from Nigeria) was also included. Transcriptomes were assembled with Trinity and proteins predicted with Transdecoder. Gene expression was estimated with FPKM mapping reads to assembled sequences, using Bowtie2 and HTSeq-count. An in-house pipeline (SignalP, TargetP, PredGPI, ProtComp, WoLFPSort and PROSITE) was used to select predicted cell-membrane and secreted proteins. The selected sequences were screened for the presence of linear B-cell epitopes with BepiPred, LBTope and IEDB tools and also scanned for intrinsically unstructured/disordered regions with IUPred. Possible cross-reactivity with other trypanosomatids was filtered with tBLASTn ($\geq 70\%$ identity over 15mer). A total of 23, 24, 3 and 68 epitopes at 18, 19, 3 and 57 proteins were selected for Til, Tsp, Tvv and Tvi, respectively. Most of the identified epitopes are present in proteins annotated as hypothetical proteins, with variable FPKM values (from only 3.4 to 611.1). Selected epitopes metrics for BepiPred, LBTope, IEDB and IUPred tools presented similar results for those observed in a control set for experimentally validated epitopes from *T. cruzi*. Epitopes clustering revealed the presence of common and exclusive sequences between the four strains. These are *in-silico* candidates with better probabilities of positive results for future experimental specie-specific serotyping and serodiagnostic tests for *T. vivax*.

Funding support:

CAPES, FAPESP, Campus France

Within and between gene variants: tracking for potential targets for populational linkage according to the metagenomic profile in a changing freshwater environment

Marcele Laux, Ricardo A. Vialle, José Miguel Ortega, Alessandra Giani

Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais

The *Microcystis* population structure of a tropical freshwater reservoir under bloom development was analyzed through metagenomic whole genome shotgun in order to find targets of interaction among population members in a changing environment. The metagenomic profile was followed over a temporal and spatial scale and pointed to a 5-fold increase in *Microcystis* assignments after the beginning of biomass increase and a similar 5-fold increase from the dam towards the upstream sampling station. Sixteen strains were assigned in the samples and *M. aeruginosa* NIES-843 was the most abundant and frequently distributed. A two-step recruitment was performed to analyze variable and conserved genetic contents of the NIES-843 sequence-discrete population. Thirteen hypervariable regions were identified, presenting lower alignment attributes and positional symmetry among the samples. A read mapping and population Single Nucleotide Variant (SNV) calling was then performed focusing on those regions, to track the gene variability within (above 99% similarity to NIES-843) and between (95-98% similarity) the mapping, since the variations at nucleotide level can be detected even among distinct populations. The variants were called according to GATK pipeline, and classified according to protein transcription impact. Only variants in coding regions were considered. The energetic metabolism seems to be under strong pressure considering the variation between populations, given that several of the variants identified are involved in regulation of the photosynthesis and respiration systems. The PSI assembly protein Ycf3, OrfB transposases family and Nudix hydrolases family were among such set of genes, which could be potentially related to the growing biomass of the *Microcystis* population over the sampling period. Similarly, the variants identified within the population were mostly related to regulation processes, noteworthy the PatA family of microbial two-component response regulators, and a transcriptional regulator Pirin-like protein, indicating potential functional keys for intraspecific dynamics. Additionally, sixty one housekeeping genes presented high impact variants, most of them with variant allele frequency of 1.0 and distinct occurrence according to each strain. We highlight the Thioredoxin gene *trxA*, essential for photosynthetic growth, which presented a generally low but increasingly depth towards the development of the bloom event, showing high impact variants in most reference genomes. The *Microcystis* population presented a mosaic-like genomic structure with distinct gene contribution among the strains. The "within" and "between" non-synonymous SNVs were mostly related to transposases and regulators, potentially associated to energetic demands and growth conditions, considering that those differentially aligned target genes were mapped mostly or even exclusively in samples from bloom events.

Draft genome of *Serratia marcescens* UENF 22-GI: a plant growth promoting bacterium isolated from vermicompost

Filipe P. Matteoli¹, Pollyanna S. Lopes², Fábio L. Olivares², Thiago M Venancio¹

¹Centro de Biociências e Biotecnologia - UENF, ²Núcleo de Desenvolvimento de Insumos

Biológicos para a Agricultura - UENF

The ongoing increase in world's population and adoption of intensive farming results in large amounts of organic waste and environmental contamination. In order to mitigate these damages sustainable approaches are being tested worldwide, many of these focused on soil management. Biological properties concerning a healthy soil are still to be uncovered, nevertheless using soil's native organisms to supply human needs in agriculture is a promising strategy. In this context vermicomposting is a widely known practice to biologically stabilize green wastes using earthworms to perform waste stabilization, promote aeration and substrate fragmentation, thereby drastically increasing the microbial activity specially in the rhizosphere. Further, vermicompost has been demonstrated to be a rich source of microbial diversity, notably plant growth promoting rhizobacteria (PGPR). These bacteria are able to promote plant growth both directly or indirectly. In this work we report the sequencing and analysis of *Serratia marcescens* (strain UENF-22GI) isolated from vermicomposted material. 21,445,242 paired-end reads were sequenced in a Illumina HiSeq 2500 platform and assembled using Velvet, resulting in a 5,001,584 Mb assembly with 3,0 Mb N50. By annotating the genome with RAST, we found 4662 genes, 30 pseudogenes, 6 rRNAs, 84 tRNAs. We did a comparative genome analysis and phylogenetics reconstruction using other publicly available *Serratia marcescens* genomes, we also performed extensive manual curation to find genes potentially responsible for plant growth-promoting properties observed in vitro. As a whole, our results indicate a biotechnological potential of *S. marcescens* in plant growth promoting products. Financial support CAPES, FAPERJ.

Variant in the *PDE4B* related to Acute Lymphoblastic Leukemia relapse is differentiated in Native Americans

Rennan G. Moreira, Fernanda Rodrigues-Soares, Eduardo Tarazona-Santos

*Laboratório de Diversidade Genética Humana. Departamento de Biologia Geral,
Universidade Federal de Minas Gerais, Belo Horizonte, Brasil*

Acute Lymphoblastic Leukemia (ALL) is the most common cancer in children, being responsible for almost 25% of the malignancies cases. Despite of recent advances in treatments, leading to more than 80% of cure rate, relapse is still common. In this sense, Pharmacogenetics is being increasingly applied on cancer cases as a new approach to improve treatment outcomes, opening new possibilities for reducing relapse events in ALL cases. Accordingly, ALL relapse risk varies considerably depending on the ethnicity. Yang et al. 2011 showed that Native American genomic component was associated with higher risk of ALL relapse in a study encompassing more than 2,500 ALL cases from the USA but also including worldwide samples for comparison. Authors identified strong association signals into the *PDE4B* (rs6683977) and *MYT1L* (rs17039396) genes in Hispanic samples. The *PDE4B* gene is involved in the metabolism of certain drugs used in cancer treatment, such as prednisone, what could suggest an effect on ALL relapse susceptibility. Thus, in order to support such findings, the main goal of this study is to evaluate genetic diversity of Native American populations from Peru and Brazil as well as of Brazilian admixed populations in the genomic regions surrounding rs6683977-*PDE4B* and rs17039396-*MYT1L*, and discuss their impact in mapping variants related to ALL relapse in additional populations with Native American ancestry. We used the BeadXpress Illumina system to genotype 76 SNPs - previously known as ancestry informative markers (AIMs) - and also 48 *PDE4B* and *MYT1L* SNPs (including the rs6683977 and rs17039396 variants) in 255 Native Americans from Peru, 88 from Brazil, and 98 admixed Brazilians. Results showed that whereas Native Peruvians have less than 5% of non-Native American ancestry, Native Brazilians present higher European and African admixture proportions. Native American ancestry proportion was inversely related to higher heterozygosity values in both genes. Population pairwise FST measures performed by each gene showed expected patterns of differentiation for human populations worldwide. However, when analyzing SNPs individually, the highest FCT value found in *PDE4B* SNPs is for rs6683977. Thus, considering allele frequencies and AMOVA analyses, we confirmed our expectation that the rs6683977-C-*PDE4B* allele is highly differentiated in Native Americans from Peru and Brazil, when comparing with Europeans. We did not see the same pattern for the ALL admixture mapping hit rs17039396 in *MYT1L*.

LTR retrotransposons in *Hemileia vastatrix* genome

Rafaela Leite Prado Rocha¹, Pedro Ricardo Rossi Marques Barreirros¹, Tiago Antônio de Oliveira Mendes², Laércio Zambolim^{1,3}, Ney Sussumu Sakiyama^{1,4}, Eveline Teixeira Caixeta^{1,5}

¹*Laboratório de Biotecnologia do Cafeeiro, BIOAGRO UFV;* ²*Departamento de Bioquímica e Biologia Molecular, UFV;* ³*Departamento de Fitopatologia, UFV;* ⁴*Departamento de Fitotecnia, UFV;* ⁵*EMBRAPA Café*

Rust is the most harmful disease that affects coffee trees, which may cause drastic drops in productivity if not controlled. The pathogen fungus *Hemileia vastatrix* displays high levels of genetic variability leading to appearance of new races and supplanting the resistance of varieties of coffee obtained in breeding programs. Since sexual reproduction was not previously related, the mechanism that causes such variability is yet not known. Thus, there is a chance that the DNA transposable elements are responsible for increasing genetic variability of the fungus. The goal of the first stage of this study was to identify the presence, frequency and location of LTR retrotransposons, a subclass of transposons present in the genome of *H. vastatrix*. For the analysis, we used an in house database of DNA sequences of *H. vastatrix* reference genome that is being developed in our group. This genome, which will be the first genome of this species, has been assembled and is available for analysis of the working group. The assembly of this genome contains 58,535 contigs with N50 and N90 of 11,385 and 3,762 bp, respectively. To search LTR retrotransposons we used LTR-Finder online software. A total of 1,117 LTR retrotransposons were found in different contigs with a density ranging from 1 to 12 LTR per contig. When we look at the density of those LTR in the genome it was found that a large number of LTR are present only in a few contigs and none or few can be found in many contigs. This indicates that there is not a pattern of insertion of those LTR retrotransposons throughout the genome. It is known that if a retrotransposon insert itself around genes that could change genes pattern of expression or function from multiple mechanisms. A mechanism example is the RIP (repeat-induced point mutation) that is a point mutations induced by repetition associated with mechanisms of defense against transposable elements. Studies suggest that such point mutations can extend beyond those repetitive regions and may lead to mutation on coding regions of the genome. Thus, the next step of this project will be to predict the proteins present in contigs where the retrotransposons were found and analyze the possible influence of those on coding regions. Take together the large number of retrotransposons present in *H. vastatrix* genome and those may be related to high genetic variability of this fungus that causes an important coffee disease.

Acknowledgements: CAPES, CNPq and FAPEMIG.

Linear Algebra Methods for Inferring Phylogenies Based on Peptides Frequencies Vectors: An Efficient Alternative Method to Investigate Relationships among Genes, Genomes and Organisms

Lara Maria Silva Miranda¹, Gabriel Bandeira Tofani¹, Gustavo Palmer Irffi¹, Lucas Felipe Silva¹, Matheus Allef Cruz¹, Thiago do Carmo Librelon Rocha¹, Bráulio Roberto Gonçalves Marinho Couto¹, Marcos Augusto dos Santos²

¹*Centro Universitário de Belo Horizonte (UniBH)*, ²*Universidade Federal de Minas Gerais (UFMG)*

The objective of this paper is to answer four questions: Is it possible to represent proteins and genomes as tripeptide frequency vectors? Why Euclidean distance between protein vectors is better than the cosine as a metric to build phylogenetic trees? Are phylogenetic trees constructed by using Euclidean distance between protein vectors consistent with phylogenetic trees constructed with alignments (classical phylogenetic trees)? Do images of genomes represented by multidimensional vectors and visualized in reduced tridimensional space generate relationships among species consistent with those described by classical phylogenetic trees? Five sets of sequences were analyzed by classical phylogenetics techniques, based on pairwise alignments, and by Linear Algebra and optimization methods. Firstly, the origin of the Human Immunodeficiency Virus (HIV) was analyzed, retrieving from GenBank the three longest coding regions from seventeen different isolated strains of the Human and Simian immunodeficiency virus (SIV). The second database was composed by the complete genome of five strains of Chlamydophila pneumoniae that were retrieved from the NCBI (National Center for Biotechnology Information) website. Wholegenome sequencing of MRSA isolates from 14 patients involved in a outbreak were the third database. The fourth dataset was composed by 59 whole mitochondrial genomes from the NCBI genome database, each one with 13 genes, totaling 767 proteins. The last database analyzed was composed by mitochondrial D-loop sequences for the Hominidae taxa (pongidae). The results showed that primary protein sequences and genomes can be represented as vectors in multidimensional space in such way that when they are mapped into 3D space the relationships among species are consistent with classic phylogenetic trees. Computationally, and mathematically the proposed method simplifies the study of the evolutionary chain of genes and genomes. The computational load is substantially lowered and complete genomes can be easily analyzed in a very modest computer.

The genomic basis for the variable biochemical profiles that lead to erroneous identifications of emerging pathogenic *Corynebacterium* spp.

André S. Santos¹, Catarina A. Moreira², Carolina S. Silva², Arthur Silva³, Vasco A. Azevedo¹, Liza F. Vilela¹, Luis G. C. Pacheco²

¹*Universidade Federal de Minas Gerais*, ²*Universidade Federal da Bahia*, ³*Universidade Federal do Pará*

Emerging and reemerging pathogenic bacteria of the genus *Corynebacterium* have been increasingly recognized as the causative agents of infections in humans. The identification of these bacteria by the most commonly used phenotypic tests, based on batteries of twenty-one biochemical reactions, is considered as challenging, and normally requires additional methods including 16S rRNA gene sequencing and MALDI-TOF mass spectrometry. In particular, the frequently reported species *C. amycolatum*, *C. striatum*, *C. xerosis* and *C. minutissimum* (XSMA group) normally generate biochemical profiles that may lead to ambiguous and even erroneous identifications in the clinical microbiology laboratory. Besides, the most frequently found species *C. diphtheriae*, which causes the reemerging disease diphtheria, may present a variable carbohydrate fermentation profile, then hampering appropriate identification by commonly used biochemical methods. In order to study the genetic basis for the variable profiles observed for these bacteria in biochemical tests, we performed a comparative genomic analysis between 13 strains of *C. diphtheriae* and between several isolates of bacteria of the XSMA group. Various reactions and metabolic pathways that show variability in biochemical tests were targeted in this work, including: sucrose, galactose, maltose, ribose and glycogen utilization, and nitrate reduction. The enzyme codes for each of metabolic reactions were obtained from MetaCyc database and used to obtain the sequences of proteins in the UniProt database. TBLASTN searches were performed using the following cut-off parameters: 70% query cover, 30% identity and expect value $< 10^{-4}$. Among the *C. diphtheriae* strains it was identified a lineage-specific presence of sucrose degradation pathway, this result provides evidence of the biochemical plasticity observed in this species, since the literature indicates as sucrose non-fermenting bacteria. Some genes necessary for the maintenance of the metabolic pathway were identified in genomic islands present in strains of *C. diphtheriae* 31A and BH8, this is a major finding to account for the typically observed variability. The proteins necessary for maintenance of the pathway of degradation of sucrose were modeled and aligned with corresponding proteins of *C. glutamicum*, a species presenting sucrose utilization capacity, and showed high similarity. The biochemical variability of the XSMA group, observed in the literature for these reactions were confirmed by genomic analysis *in silico*.

Infering the genetic structure and the history of interaction between Andean and Amazonian human populations using genome-wide data

Victor Borda^{1,2}, Marilia Scliar¹, Mateus Gouvéia¹, Thiago Leal¹, Gilderlanio Araújo¹, Giordano Soares-Souza¹, Robert H Gilman^{3,4}, Heinner Guio², Eduardo Tarazona-Santos¹

¹ Laboratorio de Diversidade Genetica Humana, Instituto de Ciencias Biologicas, Universidade Federal de Minas Gerais. ² Peruvian National Institute of Health, Ministry of Health, Perú. ³Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD 21205. ⁴Laboratorio de Investigación de Enfermedades Infecciosas, Universidad Peruana Cayetano Heredia, 15102, Perú

The peopling of the Americas implied a complex process in which human population entered into North America and migrated until the tip of Chile. In this process, the settlement of Amazon and Andean regions involved many demographic processes that have a strong influence on cultural and genetic flow. We use genome-wide information of several native populations from the Coast, Andes and Amazon Peruvian regions in order to disentangling the genetic structure and the historical relationships among them, as well as to reveal the genetic structure of biomedically relevant variants in these populations. We analyzed a dataset of 153 unrelated samples for 2.5 million of SNPs from 10 Natives Peruvian populations. This study involves collaboration between our group and the Peruvian National Institute of Health (INS). We use two different approaches to detect genetic structure: SNP and haplotype-based analyses. For SNP-based analysis we used PCA EIGENSOFT and ADMIXTURE and for haplotype-based we used CHROMOPAINTER and fineSTRUCTURE. SNP-based methods showed two groups. A first group included Andean populations (South-Central Andean Quechuas [SCAQ], Chopccas, Qeros, Aimaras and Uros) and Coastal population (Moches). The second group included the Amazon populations (Ashaninkas, Matsiguenkas, Matses and Nahuas). The haplotype-based methods showed a fine differentiation within the Andean and Amazon populations. On the Andean region, we observed the separation of Quechua-speaking populations (SCQA and Chopccas) from the Aimaran-speaking (Aimaras and Uros) except for Qeros (Quechua-speaking), that showed more affinities with the Aimaran group. Also, both approaches showed some affinities of Moches (Shore) with Matses (North Amazon) and between High Amazon (Matsiguenkas and Ashaninkas) with Aimaran speaking populations (Aimaras and Uros). We are also identifying variants that are highly differentiating between the Andes and Amazon that may be involved in adaptations to these environments, and using a network approach to define the genetic structure of GWAS-hits in these populations.

Funding support: The first author is part of the Postgraduate Program in Bioinformatics –UFMG and of the Programa Estudante Convênio Pós-Graduação of CAPES (PEC-PG). This study received financial support from CNPq and INS.

Genotypic characterization of *Vibrio parahaemolyticus* strains isolated in Brazil

Cristóvão Antunes de Lanna, Leandro Santos, Paulo Mascarello Bisch, Wanda Maria Almeida von Krüger

Laboratório de Física Biológica (FisBio), Instituto de Biofísica Carlos Chagas Filho (IBCCF), Universidade Federal do Rio de Janeiro (UFRJ)

Vibrio parahaemolyticus is a Gram-negative bacterium that inhabits marine and estuarine environments. Many of its strains are pathogenic to man. In Brazil, *V. parahaemolyticus* has caused intestinal infection outbreaks following the consumption of raw or undercooked seafood, and has been isolated from clinical and environmental samples. Its main virulence factors are two hemolysins, TDH and TRH, and type III (T3SS) and type VI (T6SS) secretion systems. However, it is possible that other factors are involved in the pathogenicity of this species. We have been studying two environmental strains of *V. parahaemolyticus* (IOC 20128/10, 20138/10 IOC), isolated from an oyster farm in Santa Catarina state. PCR analysis indicated that both strains carry *tth* gene that is a marker for the species, but do not have the TDH and TRH hemolysin genes. To be able to further characterize these isolates, compare them to reference strains from clinical and environmental origins, to identify novel and conserved features and to determine genotype-phenotype relationships, the whole genomes of IOC 20128/10 and 20138/10 IOC were sequenced using Illumina MiSeq. Following reads quality evaluation, with FastQC and Trimmomatic, the filtered reads were assembled in contigs using Velvet and Spades algorithms. Number and average size of contigs, as well as N50 values for the two assemblies were then compared.

The contigs of the draft genomes were mapped to the two chromosomes of the reference environmental strain, *V. parahaemolyticus* BB22OP, using Contiguator. The GView tool was then used to generate graphical representations of the bacterial chromosomes. We found that both strains contain two circular chromosomes of 3.1 and 1.7 Mb, with a mean G + C content of 45.6%. The assembled genomes were subjected to automated annotation using the RAST online tool, and searched for virulence-related genes. The absence of the virulence-related genes *tdh* and *trh* was confirmed. Furthermore, two clusters containing T3SS genes were identified in the genome of both strains. A conserved synteny of the T3SS cluster genes was observed by comparing the genomes of IOC 20128/10 and 20138/10 IOC with those of the reference strains, RIMD2210633 and BB22OP, and the clinical isolates, Cascavel and 17384. Genes of components of type VI secretion systems (T6SS) were also identified in the genome of IOC 20128/10 and 20138/10 IO. Further analysis of their complete genome sequences will help to evaluate their pathogenic potential, to analyze their similarity and/or differences and could provide insight into the diversity of this species.

Searching for genomic elements of sexual reproduction in a microsporidian pathogen

Juliano de Oliveira Silveira, Jean-François Pombert, Karen Luisa Haag

Universidade Federal do Rio Grande do Sul and Illinois Institute of Technology

The adaptive value of sex in eukaryotic unicellular pathogens is a matter of intense debate, and the genetic mechanisms involved in sexual reproduction of these organisms are largely unknown. Differences in reproduction systems could help explaining differences in modes of transmission, host range amplitude, and response to environmental changes, all tightly related to pathogenicity and virulence of such pathogens. This work aims to identify genomic elements that could be associated with modes of reproduction and transmission, and further elucidate the role of sexual cycles in a microsporidian parasite. Microsporidia belong to a phylum of unicellular intracellular pathogenic Fungi. These parasites are found in virtually all types of animals, being of importance to health and agriculture. *Hamiltosporidium tvaerminnensis* is a microsporidian shown to be asexual and being both vertically and horizontally transmitted to its host, the microcrustacean *Daphnia magna*. *H. tvaerminnensis* possesses a sister species, *Hamiltosporidium magnivora*, which differs by being sexual and only vertically transmitted. Our work uses as reference a newly assembled draft genome of *H. tvaerminnensis*. It is an unusually large genome for a microsporidian, estimated to contain 25 Mb of sequence. Additionally, we rely on deep sequencing reads from two recently resequenced lineages of *H. magnivora*, one from Belgium and another from Israel, as well as a set from *H. tvaerminnensis*, from Finland, using technology newer than the reference. Our sequence assembly yielded 11.87 Mb for the genome of *H. tvaerminnensis* Finland strain, with a GC content of 27.29%, and at least 10-fold reduction in contig number comparing to the reference. Assemblies of *H. magnivora* isolates from Belgium and Israel yielded 17.42 Mb and 18.03 Mb respectively, both with a GC content of 32.78%. Continuing on, we plan to use two approaches in our study: (1) Comparing overall gene contents and searching for point mutations or indels in reproduction related genes, by mapping the reads of each Hamiltosporidium isolate onto contigs containing meiosis and cell division candidate genes. We are currently annotating those genes on the assemblies. (2) Searching for larger rearrangements in the Hamiltosporidium genomes, by aligning the assembled genomes.

We acknowledge the Brazilian Coordination for the Improvement of Higher Education Personnel (CAPES) for providing funding, and Dr Dieter Ebert for supporting this work.

Identification and variability analysis of monooxygenase gene family from *Chrysoporthe cubensis*

Túlio Morgan, Murillo Peterlini Tavares, Refaela Inês de Souza Ladeira Ázar,
Tiago Antônio de Oliveira Mendes, Valéria Monteze Guimarães

Universidade Federal de Viçosa, Departamento de Bioquímica e Biologia Molecular

The increasing interest in renewable sources of energy and materials has motivated studies in biochemical conversion of lignocellulosic materials to valuable products. The enzymatic saccharification is one of the critical steps of converting lignocellulosic material, characterized by applying enzymes to break down the polymers in their basic constituents. Some limiting factors of the enzymatic hydrolysis include high cost of the enzymes and the recalcitrance of lignocellulose. Thus, the improvement of enzyme cocktails are continually sought. In this context, the auxiliary active enzymes has emerged as enhancers of efficiency of saccharification and the enzymes belonging to AA9 family, namely Lytic Polysaccharide Monooxygenases (LPMO), is one of the most promising family. Our research group has been working with the phytopathogenic fungus *Chrysoporthe cubensis*, which has been showing more efficient than some commercial enzyme cocktails for saccharification. This has motivated to search for possible factors that explain this high saccharification efficiency and of them, is possibly the secretion of AA9 enzymes. In order to evaluate this hypothesis, protein sequences belonging to AA9 family from eukaryots obtained from CAZy were aligned using tblastn algorithm with the complete draft genome of *C. cubensis* and only the alignments with more than 70% of coverage and 40% identity was selected as candidates of AA9 coding genes. It was observed 169 homologous regions and these regions were subject to a manual curation step for eliminating redundancies, resulting in 12 specific sequences. In order to obtain the complete CDS, the region coordinates were subjected to *ab initio* gene prediction using the program Augustus set to *Neurospora crassa* as model organism. Taking together, these results provide a strong indicative of AA9 production and secretion by *C. cubensis* and accounts for high potential of this fungus in biotechnological process. Moreover, these results opens the possibility of heterologous expression of these enzymes, enabling improvements in saccharification yields as well as biochemical and molecular studies of LPMOs. Acknowledgements: CAPES, CNPq and FAPEMIG.

In silico prediction of auxiliary activity enzymes secreted by the fungus *Chrysoporthe cubensis*

Murillo Peterlini Tavares¹, Túlio Morgan¹, Thiago Rodrigues Dutra¹, Tiago Antônio de Oliveira Mendes¹, Hugo Rody Vianna Silva², Valéria Monteze Guimarães¹

¹*Universidade Federal de Viçosa, Departamento de Bioquímica e Biologia Molecular, Viçosa, MG – Brazil*, ²*Universidade Federal de São Paulo, Instituto de Ciência e Tecnologia, São José dos Campos, SP – Brazil.*

The growing concern over the worldwide shortage of fossil fuels and the increasing emissions of greenhouse gases has provided the development of technologies that use biomass from agricultural residues for the production of biofuels. Our research group has demonstrated that *Chrysoporthe cubensis*, a plant pathogenic fungus, has produced enzymatic extracts more efficient for plant biomass degradation than commercial preparations. So, it is essential to know in detail the enzymes and proteins secreted by this fungus, especially those involved in the hydrolysis of biomass. In addition to the Glycoside Hydrolyses (GH) enzymes, this fungus can produce Auxiliary Activities (AAs), which are still poorly studied. The presence of this wide range of enzymes may explain the high efficiency of this fungal extract compared to commercial cocktails. It has been proposed a bioinfosecretome study of *C.cubensis* enzymes with Auxiliary Activities, through in silico predictions of candidate protein secretion using bioinformatics tools. Computational analysis will provide information on the probable secretome of *C.cubensis* and the identification of key enzymes that can be targeted to increase the hydrolytic efficiency, making this extract more interesting for commercial purposes. Protein sequences from eukaryotes belonging to thirteen families of Auxiliary Activities enzymes of Carbohydrate-Active enzymes database (CAZy) were recovered and aligned (tblastn) with the complete genome draft of *C.cubensis* selecting only the aligned regions with more than 70% of coverage and 40% of identity. The resulting gene coordinates were subjected to a manual curation step for elimination of possible redundancies and then subjected to ab initio gene prediction using the program Augustus. After obtaining the final model of predicted genes, members of AA1 families were selected for comparative studies of sequence variability. Later, the *C.cubensis* enzymes of interest will be selected and the structural modeling by comparison with structural domains of auxiliary enzymes characterized and belonging to fungi with different lifestyles, in which it will allow to verify unique characteristics of the catalytic sites of *C.cubensis* enzymes that can positively influence the interaction between target and substrate. Thus, through comparative studies of structural diversity among the members of AA1 family, is expected to select enzymes with great potential for commercial application.

Acknowledgements: CAPES, CNPq and FAPEMIG.

A novel hierarchical *in silico* approach for the prediction of drug and vaccine targets against *Chlamydophila* *pneumoniae*

Ana Carolina Barbosa Caetano; Sandeep Tiwari; Núbia Seyert, Brenda Rosa da Luz; Roselane Gonçalves Ribeiro; Thiago Luiz de Paula Castro; Vasco Azevedo.

Federal University of Minas Gerais, Belo Horizonte, MG, Brazil

Chlamydophila pneumoniae is one of the most important and well-studied gram-negative bacterial pathogens. This obligate intracellular bacterium is a major cause of pneumonia and is associated with the development of other respiratory diseases in humans, including Chronic Obstructive Pulmonary Disease (COPD), chronic asthma, pharyngitis and bronchitis. According to the World Health Organization, COPD is predicted to become the third leading cause of death by the year 2030. Although much is known about the biology of *C. pneumoniae*, particular attention should be given to the development of strategies to contain infection by this bacterium. The era of Next Generation Sequencing (NGS) is pushing forward genome-based studies for the prediction of therapeutic targets against a variety of diseases caused by pathogenic microorganisms. To date, 13 complete genome sequences of *C. pneumoniae* were made available on NCBI, allowing us to conduct the search for candidate therapeutic targets against this bacterium. For this, we performed a comparative genomic analysis using a novel hierarchical approach. In Phase I, four different sets of proteins were mined through analysis of chokepoint, pathways, virulence factors, resistance genes and protein networks. In Phase II, selected protein sets were filtered through subtractive channel analysis to find out targets that are likely to be essential for the survival of the pathogen and non-similar to proteins present in the human intestinal microbiome. Finally, in Phase III, the candidate targets were qualitatively characterized by analyzing cellular localization, broad spectrum, interactome involvement, functionality, and druggability. A total of 595 non-human homologous proteins was identified and submitted to reverse vaccinology, evaluating antigenic properties of vaccine candidates. These proteins also went through subtractive and modelomics approaches for drug target identification. Based on these analyses, we classified 36 gene products as secreted proteins, putative surface-exposed proteins or membrane proteins. By using modelomics, a total of 8 cytoplasmic proteins constituting distinct quality model were selected as putative drug targets. These proteins were subjected to virtual screening using two different compound libraries extracted from the ZINC database and plant-derived natural compounds. The proposed drug molecules exhibit favorable interactions, lowered energy values and high complementarity with the predicted protein targets. The outcome of this study constitutes a preliminary step for the development of novel strategies to combat human infections caused by *C. pneumoniae*.

Establishment of a pipeline for 16S-based metagenomic studies of *Mycobacterium leprae*

Felipe Borim Corrêa^{1,2}, Gabriel da Rocha Fernandes²

Universidade Federal de Minas Gerais, Programa Interunidades de Pós-Graduação em Bioinformática, Belo Horizonte, MG, Brazil¹, Centro de Pesquisas René Rachou, Fundação Oswaldo Cruz, Biosystems Informatics and Genomics Group, Belo Horizonte, MG, Brazil²

Identifying the pathogen *Mycobacterium leprae* is not a simple task and leprosy still remains a health problem all over the world. Sequence analysis of 16S ribosomal RNA has been used to perform metagenomic studies. However, 16S-based techniques are known to have limitations because of the biases mainly related to DNA extraction and PCR amplification. Whereas we can not simulate DNA extraction, the choice of primers is an important step since the hypervariable regions for distinguishment among taxa can vary, such as the amplification efficiency. The purpose of this study was to establish a pipeline for 16S-based metagenomic studies of *Mycobacterium leprae*. Methods were divided in two parts: candidates selection and candidates evaluation. In the first part we used Simulate_PCR to perform an in silico PCR with the 16S rRNA gene sequence of *M. leprae* TN strain (NC_002677.1). Amplicons were simulated using all viable pairings of 22 forward and 22 reverse primers and were filtered by maximum length of 550 nucleotides. Primer pairs were checked for possible cross dimerization and melting temperature range. Each selected amplicon was submitted individually to QIIME for taxonomy assignment with OTU similarity of 97%. Compatible OTU databases used were Greengenes versions 6oct2010, 29nov2010, 4feb2011 and Silva version 111. For candidates evaluation we performed an in silico PCR with the selected primers from the last step using Silva and Greengenes 16S fasta databases. In candidates selection we got a total of 18 amplicons assigned taxonomically to *Mycobacterium leprae* OTUs in at least one database. Maybe the better hypervariable regions for taxonomy assignment are V2 and V6 because only amplicons which covers these regions were assigned to *M. leprae* in all databases. In candidates evaluation we got 9 primer pairs which could amplify at least 50% of Bacteria domain and 1 primer pair of Archaea for both databases. We can conclude that there are more efficient hypervariable regions for *Mycobacterium leprae* identification in environmental samples, however in the amplification step we can have a huge loss of information. In further analysis we are going to evaluate taxonomic classification with a mock community dataset.

Impact of non-synonymous mutations in adaptive diversification and domestication of soybean

Kanhu Charan Moharana¹, Thiago Motta Venancio¹

¹*Centro de Biociências e Biotecnologia, Universidade Estadual do Norte Fluminense Darcy Ribeiro, Campos dos Goytacazes, Rio de Janeiro, Brazil*

Identification of adaptive mutations is of great interest in understanding and improving desirable phenotypes in plant genetic improvement programs. However, most DNA mutations are generally neutral or slightly deleterious. Recently genome re-sequencing and identification of many trait-associated single nucleotide polymorphisms (SNPs) in wild (*Glycine soja*) and modern (*Glycine max*) soybean cultivars have been reported. However, the adaptive potential of such SNPs remain poorly studied. In the present work we used the publicly available interspecies-specific SNP data to predict mutations with potentially beneficial impact on protein structures. In a bottom-up approach we used *G. max* reference genome (Glyma1.0) to study the impact of SNPs located within quantitative trait loci. We identified 1841 unique non-synonymous SNPs along 1212 *Glycine max* genes. The effect of such substitutions depends on the extent to which it affects the protein structure. Hence a protein structure based approach was used to estimate the free energy difference ($\Delta\Delta G = \Delta G_{G.\max} - \Delta G_{G.\text{soja}}$) caused by each amino acid substitution. Using homology modeling we created theoretical protein structures encoded by *G. max* genes. Only 247 encoded proteins (20% of 1212) showed significant hit in protein-BLAST search against the PDB database. We used the FoldX software to simulate the impact of the non-synonymous mutations in the candidate proteins. We mutated amino acid residues with that present on wild-soybean and measured the change in free energy. While 58 mutations were observed to have destabilization ($\Delta\Delta G \geq 1\text{kcal/mol}$) effect, 47 mutations were observed further enhancing the protein structure stability ($\Delta\Delta G \leq -1\text{kcal/mol}$) in modern-soybean. Annotations based on *Arabidopsis* orthology indicated that these mutations are involved in various biotic and abiotic stress resistances. For example a starch synthase gene has undergone mutations that made it more stable in cultivated soybeans. Several genes involved in disease resistance also displayed differential stability between wild and cultivated soybeans. Although preliminary, our results highlight the importance of deeper computational analyses of GWAS and next-generation sequencing data in identifying the genetic basis of agronomically relevant phenotypes.

Funding: The authors acknowledge Universidade Estadual do Norte Fluminense Darcy Ribeiro, Brazil and the following Brazilian funding agencies for supporting their research: Fundação Carlos Chagas Filho de Amparo a Pesquisa do Estado do Rio de Janeiro (FAPERJ) and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES).

Prediction and analysis of plasmids from multidrug-resistant *Klebsiella pneumoniae* and *Enterobacter aerogenes* clinical isolates

¹Hemanoel Passarelli Araujo, ¹Filipe Pereira Matteoli, ²Jussara Kasuko Palmeiro,

²Líbera Dalla-Costa, ¹Thiago Motta Venancio

¹*Universidade Estadual do Norte Fluminense*, ²*Universidade Federal do Paraná*

Bacteria typically carry extrachromosomal, self-replicating genetic elements known as plasmids. Hospital outbreaks have become increasingly prevalent due to the evolution and spread of multidrug resistance, particularly in the presence of selective antibiotic pressure, which is a permanent public health concern. Here we report a comparative genomic analysis of southern Brazilian nosocomial *Klebsiella pneumoniae* and *Enterobacter aerogenes* isolates representing different resistance profiles. We sequenced four *E. aerogenes* and six *K. pneumoniae* strains, isolated from hospital patients, through Illumina Hiseq 2000 platform. Plasmids were predicted using PlasmidSPAdes and genes annotated with RAST, Prokka, BLAST and srst2 servers. We integrated our computational analysis with results from *in vitro* antibiotic resistance tests. The results show that resistance genes *blaCTX-M15*, *blaOXA-2*, *blaTEM*, *blaKPC-2*, *blaOXY*, *blaSHV*, *aac(6')-Ib-cr*, *sul1* and *qnrB1* were variably distributed among the isolates and presumably contributed to the diverse multidrug-resistance profiles. The *blaKPC-2* genes were found to be located within Tn4401b and associated with plasmids of similar Inc groups in *K. pneumoniae*. Overall, our detailed analysis not only allowed the identification of the genetic basis of the observed antibiotic resistance phenotypes of these ten strains, but also revealed potential novel resistance to antibiotics that have not been tested yet.

Chromosomal copy number variation reveals extensive levels of genomic plasticity among and within *Trypanosoma cruzi* DTUs

João Luís Reis-Cunha¹, Gabriela F. Rodrigues-Luiz¹, Hugo O. Valdivia², Rodrigo P. Baptista³, Laila Viana de Almeida¹, Mariana Santos Cardoso¹, Tiago A. O. Mendes⁴, Andrea M. Macedo¹, Ana Tereza Vasconcelos⁵, Gustavo Coutinho Cerqueira⁶, Daniella C. Bartholomeu¹

1-Universidade Federal de Minas Gerais, 2-U.S Naval Medical Research, 3-The University Of Georgia, 4-Universidade Federal de Viçosa, 5-Laboratório Nacional de Computação Científica, 6-Broad Institute.

The taxon *T. cruzi* is divided into six discrete typing units (DTUs), named TcI-TcVI. CL Brener, the reference strain of *T. cruzi* genome project belongs to the hybrid DTU TcVI, presenting 41 putative chromosomes. Chromosomal Copy Number Variation (CCNV) is a mechanism of gene expansion possibly related to rapid adaptation to new environments, and is already documented in yeast and several *Leishmania* species. Although studies point toward karyotype variability in *T. cruzi* strains, the extent of diversity in CCNV among and within DTUs based on read depth coverage (RDC) analysis has not been determined. To identify CCNV among *T. cruzi* DTUs, we sequenced genomes of strains from TcI, TcII and TcIII DTUs and estimated the ploidy based on RDC of single copy genes in each chromosome. TcI strains had few aneuploidies, while strains from TcII and TcIII DTUs presented a high degree of chromosomal expansions, which is in agreement with the average DNA mass per cell and genome plasticity in these DTUs. Chromosome 31, the only supernumerary chromosome in all *T. cruzi* samples evaluated, is enriched with genes related to glycosylation pathways, such as the enzyme UDP-GlcNAc-dependent glycosyltransferase, involved in the initial steps of mucin glycosylation. As the strains from the TcII DTU presented a divergent pattern of chromosomal expansions, we sequenced the genome of 7 *T. cruzi* TcII field isolates from Minas Gerais state, Brazil. These samples presented a complex pattern of chromosomal duplication/loss, which is not in agreement with the phylogeny based on single copy genes. Finally, we sequenced three clones of the TcII Y strain, which presented the same CCNVs as the non-cloned population, suggesting stability in the chromosomal expansions/loss pattern in the population of Y strain. Increased gene copy number due to chromosome amplification may contribute to alterations in gene expression, representing a crucial strategy for parasites that mainly depend on post-transcriptional mechanisms to control gene expression.

Modular variability of multigene families encoding surface proteins uncovers differential composition of motifs among *Trypanosoma cruzi* strains

João Luís Reis-Cunha¹, Gabriela F. Rodrigues-Luiz¹, Rodrigo P. Baptista², Laila Viana de Almeida¹, Mariana Santos Cardoso¹, Gustavo Coutinho Cerqueira³, Daniella C. Bartholomeu¹

1-Universidade Federal de Minas Gerais, 2-The University of Georgia, 3-Broad Institute.

Among the Tritryps, *T. cruzi* owns the largest expansion of multigene families encoding surface proteins. Despite playing crucial role in host-parasite interactions, one third of these gene families were not incorporated into the 41 putative chromosomes in the *T.cruzi* reference strain CL Brener. The large number of members of these families also hinders the assignment of reads to a specific gene, as they can map/align with the same reliability to several loci. Although these families are highly polymorphic, they also present motifs shared among distinct members, resulting in a mosaic structure that may favor the generation of sequence variability by rearrangement of defined blocks through recombination. The relative abundance of these conserved motifs can be used to estimate the variability of these regions among *T.cruzi* strains. To this end, we developed a methodology to evaluate the copy number variation of motifs derived from mucin-associated surface protein (MASP), TcMUC mucins and trans-sialidases multigene families. This methodology is assembly independent and only requires next generation sequence reads for a given isolate and a reference genome. The first step of this methodology consists in retrieving all reads that map with all the genes of each family, generating all possible kmers of 30 nucleotides present in these reads. The kmers are then clustered by sequence similarity to generate conserved motifs. Finally, the deep of coverage of each motif is computed and compared among *T. cruzi* strains. Our methodology was used to estimate the relative abundance of all motifs identified in MASP, mucin and trans-sialidase families in different *T. cruzi* DTUs, revealing several differences in their abundance within and among DTUs. Dendograms based on the abundance of these motifs presented discordances with the phylogeny based on single copy genes, reinforcing the hypothesis that different selective pressures shape the evolution of these two *T. cruzi* genomic regions.

Genomic identification and patterns of expression of secondary metabolite gene clusters in the entomopathogen fungus *Metarhizium anisopliae*

Nicolau Sbaraini^{1,2}, Rafael Lucas Muniz Guedes^{1,3}, Fábio Carrer Andreis^{1,2}, Ângela Junges^{1,2}, Guilherme Loss de Moraes^{1,2,3}, Marilene Henning Vainstein^{1,2}, Ana Tereza Ribeiro de Vasconcelos^{1,3}, Augusto Schrank^{1,2}.

¹ Rede Avançada em Biologia Computacional, Petrópolis, RJ, Brazil., ² Centro de Biotecnologia, Programa de Pós-graduação em Biologia Celular e Molecular, UFRGS, Porto Alegre, RS, Brazil., ³ Laboratório Nacional de Computação Científica, Petrópolis, RJ, Brazil.

The *Metarhizium* genus harbors cosmopolitan fungi that infect arthropod hosts. Importantly, while some species infect a wide range of hosts (host-generalists), other species infect only a few arthropods (host-specialists). This singular evolutionary trait permits unique comparisons to determine how pathogens and virulence determinants emerge and evolved. Among the several virulence determinants that have been described, secondary metabolites (SMs) are suggested to play essential roles during fungal infection. Nevertheless, genes related to SM production in *Metarhizium* spp. are scarcely described and little is known about their genomic organization, expression, regulation and role during host infection. Here, we have performed a deep survey and description of SM biosynthetic gene clusters (BGCs) in *M. anisopliae* and assessed conservation among the *Metarhizium* genus. RNA-seq data from fungi grown on cattle-tick cuticles (mimicking infection) was analyzed to validate some of the predictions and to access the differential expression of BGCs. Furthermore, our analysis extended to the construction of a phylogeny for the following three BGCs: a tropolone/citrinin-related compound, a pseurotin-related compound, and a putative helvolic acid. Among 73 BGCs identified in *M. anisopliae*, 20% were up-regulated during initial tick cuticle infection and presumably possess virulence-related roles. These up-regulated BGCs include known clusters, such as destruxin, NG39x and ferricrocin, together with putative helvolic acid and pseurotin- and tropolone/citrinin-related compound clusters as well as uncharacterized clusters. Concerning host-range several up-regulated BGCs were not conserved in host-specialist species from the *Metarhizium* genus, indicating possible differences in the metabolic strategies employed by generalist and specialist species to overcome and kill their hosts. These differences in metabolic potential may have been partially shaped by horizontal gene transfer events, as shown in our phylogenetic analysis. In conclusion, several unknown BGCs are described, and their organization, regulation and origin are discussed, providing support for the impact of SM on the *Metarhizium* genus lifestyle and infection process.

Genome analysis of *E. nigrum* and other filamentous fungi reveals molecular mechanisms related to endophytic/pathogenic lifestyles

Almir J. Ferreira^{1,2}, Liliane S. Oliveira², João M. P. Alves², Michael Thon³, Alan M. Durham⁴, Léia C. L. Fávaro¹, Arthur Gruber^{2*} and Welington L. Araújo^{1*}

¹Dept. of Microbiology and ²Dept. of Parasitology, Institute of Biomedical Sciences, USP, São Paulo, Brazil;

³Instituto Hispano-Luso de Investigaciones Agrarias (CIALE), University of Salamanca, Spain. ⁴Dept. of Computer Sciences, Institute of Mathematics and Statistics, USP, São Paulo, Brazil. *Correspondence:

argruber@usp.br and wlaraujo@usp.br

Proper knowledge of the genomes from the endophytic fungus *Epicoccum nigrum* and plants may contribute to improve agricultural production. The molecular mechanisms determining the pathogenic and endophytic lifestyles of fungi are not fully understood. *E. nigrum* is an endophytic fungus that has been used for plant pathogen biocontrol in different host plants, since it produces a series of secondary metabolites of biotechnological interest, including antimicrobials. We have previously determined that the *E. nigrum* isolate P16 produces secondary metabolites with antimicrobial activity. In this work, we confirm this activity by specific assays and report the genome sequencing and annotation of this isolate. We sequenced the whole genome and the transcriptome of the mycelium using the 454 and SOLiD platforms, respectively. The genome was assembled and gene prediction was performed with the MAKER package, using a dataset of Dothideomycetes proteins and *E. nigrum* transcript sequences as evidence. The genome sequence and gene predictions were submitted to a comprehensive functional annotation pipeline using the EGene2 platform. In order to identify gene clusters associated with secondary metabolites and compare their occurrence across endophytic and pathogenic fungi, we analyzed our genome sequence, together with the genomes from 12 different fungi using the AntiSmash server. Finally, the results were analyzed with Synteny Clusters, a specific tool developed by our group that compares gene clusters associated with secondary metabolite biosynthesis in different organisms. We found 10,320 protein coding genes. Based on structural features, *E. nigrum* presents a genome very similar to closely related filamentous fungi, including some with distinct lifestyles. We identified a total of 38 secondary metabolite gene clusters. The comparative analysis across different fungi revealed three clusters restricted to most of the pathogenic fungi, but absent in *E. nigrum*. These clusters are related to plant diseases and antibiotic activity and may be part of the gene repertoire required for a pathogenic lifestyle. In fact, data from the literature seems to corroborate the importance of some of these genes in pathogenicity. This result suggests that the lifestyle differences observed between endophytic and pathogenic fungi might rely on a relatively low number of genes. This conclusion is in agreement with a phylogenetic analysis using seven protein sequences from endophytic and pathogenic fungi, which revealed a close relationship between *E. nigrum*, an endophyte, and *Didymella exigua*, a phytopathogenic fungus. Support: FAPESP, CNPq and CAPES.

Combining profile Hidden Markov Models and small RNA pattern based strategies to identify novel Endogenous Viral Elements (EVEs) and exogenous viruses

Eric R. G. R. Aguiar^{1*}, Liliane S. Oliveira², João M. P. Alves², João Trindade Marques¹ and Arthur Gruber^{2*}

¹Department of Biochemistry and Immunology, Institute of Biological Sciences, UFMG, Belo Horizonte MG, Brazil; ²Department of Parasitology, Institute of Biomedical Sciences, USP, São Paulo SP, Brazil.

*Correspondence: ericgdp@gmail.com and argruber@usp.br

Endogenous Viral Elements (EVEs) are presumably derived from ancestral viruses that used to infect their hosts and had their sequences integrated into the host genomes. These elements show considerable sequence similarity to extant viral genomes. Thus, the accurate identification and characterization of novel EVEs are fundamental for the correct discrimination from exogenous viral sequences in metagenomic studies. In addition, the precise identification of EVEs enables a wide survey of ancestral viruses to which the host has been exposed (paleovirology), allowing a comparison to the viruses currently circulating in the host. Here we present a strategy to detect and discriminate EVEs and exogenous viral sequences using profile Hidden Markov Models (pHMMs) and an experimental validation using small RNA deep sequencing. We used virDB-Pack (see abstract by Oliveira *et al.* – X-Meeting 2016) to select a subset of 506 pHMMs from the vFam database according to virus-specific annotation terms. This set was used to screen a dataset composed of long RNAs sequenced from *Aedes aegypti*. Selected pHMMs showing the highest numbers of significant positive hits were employed as seeds for progressive assembly using GenSeed-HMM. The reconstructed sequences were submitted to similarity searches against the *nr* database, matching a wide variety of viral families. We have previously shown that EVEs present small RNA profiles with molecular characteristics distinct from exogenous viral sequences. Therefore, we mapped reads from small RNA libraries prepared from the same *A. aegypti* mosquitoes onto our candidate contigs to identify signatures that discriminate canonical EVEs from viruses. Notably, EVEs have small RNA profiles that show size in between 24-29 nt, U enrichment at the 1st nt of antisense reads, an enrichment at the 10th of sense reads, 10-nt overlap between 5' end of reads in opposite strands and asymmetrical small RNA density along sequence. Seventy-seven out of 381 contigs greater than 200 nt showed profiles consistent with EVEs, are not present on the current genome of *A. aegypti*, probably representing novel elements. We also found a 9-kb contig that showed a viral signature and represents a *Phasi Charoen Like-virus*, recently described by our group using small RNA-based strategy. This result confirms that both strategies, (1) pHMM screening followed by progressive assembly and (2) identification of viral signatures using small RNAs, can be jointly used to reliably identify and characterize new EVEs, even in the absence of a well-curated genome, and also to detect novel exogenous viruses. Support: CNPq and CAPES.

Preliminary analysis of functional SNPs mining from GBS data using GATK in a sugarcane map population

Alexandre H. Aono¹, Estela A. Costa¹, Hugo V. S. Rody¹, James S. Nagai¹, Anete P. de Souza^{2,3}, Reginaldo M. Kuroshu¹

¹*Instituto de Ciência e Tecnologia, Universidade Federal de São Paulo, São José dos Campos, SP - Brazil*, ²*Molecular Biology and Genetic Engineering Center (CBMEG) – University of Campinas (UNICAMP), Campinas, SP – Brazil*, ³*Vegetal Biology Department, Institute of Biology, University of Campinas (UNICAMP), Campinas - SP, Brazil.*

Sugarcane is the source of sugar in all tropical and subtropical countries and it is becoming increasingly important for bio-based fuels. However, its large (10 Gb), polyploid, complex genome has hindered genome based breeding efforts. Currently, genotyping-by-sequencing (GBS) has been the most economical approach for generating population genomic data without the need of a reference genome. Here, GBS was carried out in 182 full-sibs derived from a sugarcane commercial cross (IACSP96-3018 x IACSP93-3046) in order to establish a pipeline to identify SNPs in polyploidy species and generate informative molecular markers. After a sequencing using the platform Illumina GAIIX (1x120bp), we processed the data following the GATK pipeline, modifying it and creating in-house scripts to handle polyploidy genomes. As a first step in the pre-processing phase of the analysis, we used FASTX-Toolkit for demultiplexing and barcode processing. A comparative alignment was performed using BWA-MEM algorithm against three different references: sorghum genome (1), sugarcane RNA-seq data (2) and sugarcane methyl-filtered genome (3). Picard tools were used to mark alignment duplicates and SAMtools for controlling the process. Genotype calls were first made in gVCF format for each sample using HaplotypeCaller with stringent parameters for phred-scaled confidence threshold and ploidy level as 12. Then, all samples were joined into a VCF file as implemented in GATK 3.6 pipeline. As a result, from ~174 million reads generated, it was obtained: 94% of correspondence in (1), 33% in (2) and 41% in (3). In order to identify functional SNPs, we selected a set of aligned contigs to (2). These contigs were previously classified as part of two representative pathways: Carbon Fixation in Photosynthetic (C4 photosynthetic pathway) and the Starch and Sucrose Metabolism. From 53 selected contigs, we obtained 129 putative SNPs in 28 contigs. In C4 photosynthetic pathway, 91 putative SNPs were found and 38 in the Starch and Sucrose pathway. With these preliminary analyses we identified SNPs that can be used as candidates for the development of functional specific markers and started a process to establish a pipeline for searching SNPs in polyploidy and aneuploidy species as sugarcane. As future issues, we expect to find more SNPs using the other references for sugarcane, by mapping more and different regions; including non-coding regions.

GBS data from a population of modern sugarcane variety reveals duplicate genes retained by breeding process

Hugo V. S. Rody¹, James S. Nagai¹, Estela A. Costa¹, Alexandre H. Aono¹, Anete P. de Souza^{2,3}, Reginaldo M. Kuroshu¹

¹*Instituto de Ciéncia e Tecnologia, Universidade Federal de São Paulo, São José dos Campos, SP-Brazil*, ²*Molecular Biology and Genetic Engineering Center (CBMEG) - University of Campinas (UNICAMP), Campinas, SP-Brazil*, ³*Vegetal Biology Department, Institute of Biology, University of Campinas (UNICAMP), Campinas-SP, Brazil*

Sugarcane is the most important crop for sugar and biofuel production. Despite all economic interest, sugarcane breeding is challenging due to its overcomplex genetics, with cultivars varying in chromosome number from 80 to 130. New sequencing methods such as genotyping-by-sequencing (GBS) have accelerated studies using genomic data from populations of non-model organisms. However, because short reads are likely to map with equal probability in multiple positions, duplicate genes have been typically filtered from GBS data. We used a pipeline to expose duplicate genes in GBS data from a population of modern sugarcane variety from the Sugarcane Breeding Program at IAC/Apta, obtained using IACSP96-3046 and IACSP95-3018 as parents. Additionally, we investigated which duplicate gene categories are consistently overrepresented across the population. After GBS raw reads manipulation, final high quality reads, with minimum of 80% of Q > 20 and 85bp long, were used for De novo assembly; performed by Stacks v.1.42. To filter young duplicate genes that typically are merged as a single locus in GBS data, we allowed maximum of two nucleotides mismatches among reads to form a putative locus. Old duplicates are expected to have accumulated enough mutations to form a new locus. Using BLASTn, all consensus loci were compared to the Sorghum Coding-DNA sequence (CDS) genome, with a cutoff e-20 and minimum alignment length equal to 60. With BLASTn result, four subsets were created, grouping the genes of sorghum by the number of individuals in the sugarcane population that harbored at least one consensus locus showing similarity to respective sorghum gene. In Subset1, sorghum genes were present in a unique individual across the sugarcane population. Subset2 was formed by sorghum genes that occurred from 2 to 50 individuals, Subset3 from 51 to 100 individuals, and Subset4 by genes that occurred above 100 individuals. Gene Ontology (GO) enrichment analysis, based on sorghum annotation, was carried out for each subset. Different subsets had different GO categories overrepresented. Subset1 is enriched by gene categories whose products likely stand alone in metabolic pathways, such as “cellular response to stress”. Whereas in Subset4, most of genes overrepresented are connected genes, such as those involved with signaling. Further, essential genes for carbon fixation in C4 organisms were overrepresented in Subset4. We showed overrepresented duplicate gene categories highly and lowly distributed across the sugarcane population, suggesting how breeding process has influenced duplicate gene retention that formed the characteristics of modern sugarcane.

Characterization of phage sequences on *Corynebacterium pseudotuberculosis* genomes

Flavia Figueira Aburjaile, Amália Raiana F. Lobato, Luís Carlos Guimarães, Ana Lídia Queiroz Cavalcante, Kenny da Costa Pinheiro, Adonney Allan de Oliveira Veras, Rafael Azevedo Baraúna, Artur Silva, Rommel Thiago Jucá Ramos

Institute of Biological Sciences, Federal University Pará, Belém, Pará, Brazil

The phage infection in bacterial genomes is a process that can lead to expression atypical characteristics, such as, virulence or resistance to antimicrobials. The detection of phage on pathogenic bacteria, such as, *Corynebacterium pseudotuberculosis*, may explain the virulence phenotype expression and the adaptation mechanisms used to survive in different hosts. *C. pseudotuberculosis* is a pathogenic bacterium that affects cattle and it is classified on biovars *equi* (positive nitrate reductase) and *ovis* (negative nitrate reductase) which cause expressive losses in the livestock. 17 sequences of phage were found in *C. pseudotuberculosis* genomes. The *nrdF2* gene was conserved in all 35 strains analyzed, the *ychF* gene is conserved in all strains of biovar *ovis* and some strains of biovar *equi*. In addition, hypothetical protein (405 nc) is present in some strains of biovar *equi* and 8 sequences were conserved in all strains of biovar *ovis*. These sequences presented synteny, and some of them are differentially expressed under conditions of abiotic stresses. The *nrdF2* gene present in all the strains and related to phages is a precise marker to differentiate the biovars of bacteria by phylogeny performed. This study showed that the phage sequences were highly conserved among strains of *C. pseudotuberculosis* and, one being possible to be phylogenetic marker to differentiate the biovars. The differential expression of these sequences in response to abiotic stress and characterization of products *in silico* indicates the importance of the role of those sequences in the survival of the bacteria. However, many of these sequences still need to be annotated through functional prediction to products or homologous domains in databases, the hypothetical protein with 228 pair bases was induced in all the stresses applied *in vitro* and is conserved in all strains of biovar *ovis*, it is a potential target to be investigated in further studies.

Funding: CNPq, CAPES and PROPESP.

Comparative genomic analysis of clinical and environmental strains of *Vibrio parahaemolyticus* isolated in Brazil: insight into their virulence potential

Leandro de Oliveira Santos, Paulo Mascarello Bisch, Wanda Maria Almeida von Krüger

Laboratório de Física-Biológica, Instituto de Biofísica Carlos Chagas Filho - UFRJ

Vibrio parahaemolyticus is a Gram-negative bacteria found in marine and estuarine environments. Some strains are human pathogens and can cause diarrhea, wound infections and septicemia. *V. parahaemolyticus* have been isolated in Brazil during outbreaks of gastroenteritis, from oyster farms, fish markets and restaurants, but their genomes have not been sequenced. The major factors involved in the pathogenesis of *V. parahaemolyticus* are the hemolysins (TDH and TRH) that form pores in the host membrane causing water and electrolytes efflux, and the Type 3 (T3SS) and Type 6 (T6SS) Secretion System that inject bacterial effectors directly into the host cytosol. The gene encoding the thermolabile hemolysin (*tlh*) is considered a signature molecular marker for the species. In this work we sequenced, assembled and annotated the genomes of 5 Brazilian isolates of *V. parahaemolyticus*: 3 (Cascavel, 17381 and 17384) from fecal samples of gastroenteritis patients and 2 (20173 and 20142) of oysters from an oyster farm. Genome DNA samples were prepared from overnight cells grown in LB medium at 37°C, using the Wizard Genomic DNA Purification Kit (Promega). Integrity and concentration of the samples were evaluated by agarose gel electrophoresis. The DNA was fragmented and the paired-end sequencing libraries were prepared using the Nextera DNA Library Kit. Sequencing was performed on the Illumina MiSeq, producing 75pb paired-end sequence data. Quality of the reads was evaluated by the FastQC program. Trimming and filtering were done (Phred quality score > 20) by the Trimmomatic algorithm. The filtered reads were assembled into contigs using Velvet Optimizer and Spades, and then mapped to the two chromosome of the clinical strain, *V. parahaemolyticus* RIMD2210633, with CONTIGuator. The GView tool was used to generate graphical representations of the chromosomes. The genomes were annotated with RAST server. All strains contain two circular chromosomes of approximately 3.1 and 1.7 Mbp, with a mean GC content of 45.2%. The *tlh* gene was identified in all the 5 genomes, in agreement to results of PCR analyses. Two copies of *tdh* were identified in the genomes of the strains Cascavel and 17381, and one in the 17384 genome. The *trh* gene wasn't detected in the 5 genomes. Moreover, two clusters of T3SS genes, whose sequences slightly differ from those of the reference genome, were detected in the 5 genomes. T6SS genes were also found in all genomes. These results demonstrate that the Brazilian strains, regardless of origin, have virulence-related genes and pathogenic potential.

Identification of Pho regulon genes and Pho box-like sequences in genomes of clinical and environmental isolates of *Vibrio parahaemolyticus* from Brazil

Leandro de Oliveira Santos, Cristóvão Antunes de Lanna, Paulo Mascarello Bisch,

Wanda Maria Almeida von Krüger

Laboratório de Física-Biológica, Instituto de Biofísica Carlos Chagas Filho - UFRJ

Vibrio parahaemolyticus is a Gram-negative bacteria present mostly in marine and estuarine environments. It is a worldwide cause of food-borne gastroenteritis, associated with raw or undercooked seafood consumption. *V. parahaemolyticus* infection symptoms includes abdominal cramps, diarrhea (in certain cases with presence of blood and mucus), nausea, vomit, low fever and headache. *V. parahaemolyticus* has also been associated with wound infection and septicemia. Several strains of *V. parahaemolyticus* have been isolated in Brazilian territory during outbreaks of gastroenteritis, from oyster farms, fish markets and restaurants, but their genomes have not been sequenced. Aquatic environments are poor in inorganic phosphate (Pi), an essential nutrient for cells. In bacteria the PhoB/PhoR two-component system plays an important role in detecting and responding to the changes of the environmental Pi concentration. PhoR, a transmembrane sensor protein, has a histidine kinase activity and PhoB is a cytoplasmic response regulator that binds to DNA sequences upstream the genes (Pho boxes) and regulates transcription. The set of genes regulated by PhoB/PhoR comprises the Pho regulon. In this work, we searched for Pho regulon genes and upstream Pho box-binding sites on genomes of seven *V. parahaemolyticus* strains from clinical and environmental origins isolated in Brazil. The seven genomes used were sequenced, assembled and annotated by our group. Many putative PhoB/PhoR regulated genes were identified in the genomes of the seven *V. parahaemolyticus* strains among those annotated by the RAST server. Their sequences were then aligned to orthologous gene sequences from *V. parahaemolyticus* strains and other species, to assess sequence diversity, size and arrangement, using Clustal and MEGA7 softwares. For Pho boxes identification we used a consensus (Daniel Costa Leite, personal communication) and MEME/MAST programs. As a result, many Pho regulon gene sequences were identified in all seven genomes, such as *vp2163* for the alkaline phosphatase, *vp0569* for the response regulator PhoB, *vpa0670* for the Pi sensor PhoR, *vpa1461* for the periplasmic Pi transporter PstS, *vp0572* for the exopolyphosphatase PpX, *vp0573* for the polyphosphatase kinase PpK, and *vpa0526* for a putative anionic porin. For most of them, we observed conservation of sequence, size and arrangement in comparison to the reference strains. We also identified Pho box-like sequences upstream those genes in all the seven genomes. These results demonstrate a high conservation of the Pho regulon genes among *V. parahaemolyticus* and other species.

Identification of genetic variations in engineered yeast for xylose consumption and acetic acid resistance applied to second generation ethanol production

Sheila Tiemi Nagamatsu¹, Luige Armando Llerena Calderon^{1,2}, Lucas Parreiras^{1,2}, Bruna Tatsue^{1,2}, Angelica Martins Gomes², Gonçalo Amarante Guimarães Pereira¹, Marcelo Falsarella Carazzolle¹

¹*Instituto de Biologia - UNICAMP*, ²*Biocelere AgroIndustrial LTDA*

The second-generation ethanol is a new and promising technology that can dramatically reduce the costs and increase the production. But, while the first-generation is based on fermentable sugar (glucose, fructose and sucrose) from sugarcane using industrial yeast, the second is based on hydrolyzed biomass consisting of the residual non-food crops such as leaves, stems, grass, etc. The biomass deconstruction process generates glucose, non-fermentable sugars (mainly xylose) and inhibitors of yeast growth (acetic acid, furfural and HMF). In order to increase the yield and productivity of yeast in the second generating process, it is necessary to identify industrial robust yeast for growth inhibitors, perform genetic modifications to allow the xylose consumption by insertion of endogenous xylose pathway genes and make use of evolutionary engineering approach to improve some characteristics by several rounds of cell growth and recycling on selective growth media. In this study, genetically modified industrial yeast for xylose-consumption was submitted for three rounds of evolutionary engineering using xylose as carbon source and adding acetic acid in the last round. For each round, two evolved strains were isolated and inoculated in the next one. In a total of seven strains (including parental) were submitted for genome sequencing and bioinformatics analysis to identify all mutations in the evolved strains in comparison with parental genome. For that was necessary to develop a set of bioinformatics analysis: (1) parental genome was assembled and submitted for gene prediction and annotation; (2) sequenced reads from evolved strains were aligned into parental genome allowing mismatches; (3) copy number variation (CNV) analysis using aligned reads and Poisson distribution (cn.MOPS); (4) SNP/Indel calling using the combination of GATK and Freebayes; (5) SNP/Indel annotation using Variant Effect Predictor (VEP). The CNV analysis revealed an increase of xylose pathway genes over evolutionary timeline which can explain partially the xylose consumption profile over three evolution rounds. Moreover, the mutation analysis identified several non-synonymous SNPs distributed over the rounds contributing to a better understanding of metabolic bottleneck of xylose consumption and acetic acid tolerance in industrial yeast.

In silico approaches to predict the impact of leukemic DNMT3a mutations & Identification of leads based on drug decitabine using Complex Based Pharmacophore Mapping and Virtual Screening

Syed Babar Jamal^{1*}, Matheus Filgueira Bezerra^{2*}, Sandeep Tiwari¹, Flavia Aburjaile^{1,3}, Vasco Azevedo¹, Artur Silva³, Cintia Renata Costa Rocha^{2,5}, Marcos André Calvancanti Bezerra⁴, Antonio Roberto Lucena-Araujo⁴, Eduardo Isidoro Carneiro Beltrão^{2,5}

*co-authorship of this work

¹Laboratory of Cellular and Molecular Genetics, UFMG; ²Keizo Asami Immunopathology Laboratory, UFPE; ³Center of Genomics and System Biology, UFGA; ⁴Biological Science Center, Genetics Department, UFPE; ⁵Biological Science Center, Biochemistry Department, UFPE

DNA methyltransferases are a group of enzymes that catalyzes the addition of a methyl group in cytosines from DNA strand and are considered important regulators of differential gene expression. Recently, the next-generation technologies revealed a high frequency (21-25%) of mutations in the gene coding for DNA methyltransferase 3a (*DNMT3a*) in patients with acute myeloid leukemia (AML). These mutations disrupt the pattern of DNA methylation and conduct hematopoietic cells into malignant transformation. Despite massive efforts to better understand the functional aspects of *DNMT3a*, experimental studies are limited to only few frequent mutations. The aberrant pattern of DNA methylation in AML led to the use of *DNMT3a* inhibitors as an alternative treatment. Decitabine is a cytosine analog that binds to the catalytic domain of *DNMT3a* with an inhibitory effect and is used in AML therapy. Although decitabine is a known *DNMT3a* inhibitor, there is a lack of information concerning the interaction of decitabine with mutants *DNMT3a*. Considering the high diversity of mutations described in *DNMT3a*, we aimed to use *in silico* approaches to predict the impact of these mutations on protein function patients outcome and interaction with decitabine. For this, we selected 24 more frequent *DNMT3a* missense mutations described in databases: (COSMIC and TCGA). To predict the impact on protein function, we used the combination of six distinct tools. Based on the best scoring system we found that out of 24 evaluated mutations, 11 were classified as damaging (R882H, R882P, R882S, R803S, D781G, R792H, R736C, R729Q, S714C, G543C, C497Y), 10 as intermediaries (F909C, R882C, R882L, M880V, K841Q, K829R, R736H, R729W, P718L, G646V) and 3 as probably benign (A741V, N501T, K468R). To predict whether the score of the mutations may influence clinical outcome, we used the The Cancer Genome Atlas (TCGA) AML cohort. Of the 190 patients were included, 45 had mutations in *DNMT3a* (24%). We found that AML patients with mutated *DNMT3a* had a minor overall survival ($p=0,005$). However, when comparing different mutations grouped by the score, there was no significant difference in survival ($p=0,214$) among them. Although we could not demonstrate any difference in clinical outcome, our analysis suggests a significant biological heterogeneity in *DNMT3a* variants. Additionally, we hypothesize that the change in the enzyme structure caused by mutations may affect the drug-enzyme interaction and consequently, predict the clinical response to decitabine. Based on aforementioned hypothesis, this work reports the complex-based pharmacophore modeling to find the important pharmacophoric features essential for the inhibition of *DNMT3a* activity by virtual screening, drug-likeness predictions, protein-ligands binding interactions, binding affinity predictions and binding energy calculations.

In-silico analyses for the discovery of drug and vaccine targets in *Corynebacterium camporealensis*: A Novel Hierarchical Approach

Syed Babar Jamal¹, Sandeep Tiwari¹, Arun Kumar Jaiwal¹, Daniela Arruda Costa¹, Nilson AR Coimbra¹, Doglas Parise¹, Henrique CP Figueiredo³, Debmalya Barh⁴, Artur Silva², Vasco AC Azevedo^{1*}

¹PG program in Bioinformatics (LGCM), Institute of Biological Sciences, Federal University of Minas Gerais, Belo Horizonte, MG, Brazil. ² Institute of Biologic Sciences, Federal University of Para, Belém, PA, Brazil. ³AQUACEN, National Reference Laboratory for Aquatic Animal Diseases, Ministry of Fisheries and Aquaculture, Federal University of Minas Gerais, Belo Horizonte, MG, Brazil. ⁴Centre for Genomics and Applied Gene Technology, Institute of Integrative Omics and Applied Biotechnology, Nonakuri, Purba Medinipur, West Bengal-721172, India

The genus *Corynebacterium* contains some bacterial species having clinical and biotechnological importance. In the last 2 decades, the taxonomy characterization of this noteworthy bacterial group has improved a lot. Pathogen genome sequencing and comparative genomics have resulted in identification of large number of effector genes shown to be responsible for promoting pathogenesis in human, animal and plants. We are reporting here *Corynebacterium camporealensis* strain CIP105508. It is a gram positive, non-spore forming, non-motile and pleomorphic rod shaped bacterium that occurs singly or are arranged in palisades or v-shaped forms. The bacterium was isolated from sheep milk affected by subclinical mastitis. Furthermore, we identify effector genes clustered in pathogenicity islands (PAIs) by scanning the genome regions for atypical GC content, codon usage biased approaches and other nucleotides statistical analysis. The present study aims at identification and qualitative characterization of promising drug targets in *C. camporealensis* using a novel hierarchical *in silico* approach, encompassing three phases of analyses. In phase I, four sets of proteins were mined through chokepoint, pathway, virulence factors, and resistance genes and protein network analysis. These were filtered in phase II, in order to find out promising drug target candidates through subtractive channel of analysis. The analysis resulted in therapeutic candidates, which are likely to be essential for the survival of the pathogen and non-homologous to host. Finally, in phase III, the candidate targets were qualitatively characterized through cellular localization, broad spectrum, interactome, functionality, and druggability analysis. The study explained their subcellular location identifying drug/vaccine targets, possibility of being broad spectrum target candidate, functional association with metabolically interacting proteins, cellular function (if hypothetical), and finally, druggable property. Outcome of this study could facilitate the identification of novel antibacterial agents for better treatment of *C. camporealensis* infections.

Genotype imputation of Hereford and Bradford bovine breeds from Brazil

Maurício de Alvarenga Mudadu, Henry Gomes de Carvalho, Marcos Jun Iti Yokoo, Fernando Flores Cardoso

EMBRAPA (Informática Agropecuária, Pecuária Sul)

The bovine breeding programs in Brazil are trying to adopt the use of genetic markers, a procedure called genomic selection (GS). GS consists in genotyping a given reference population with known phenotype and in the discovery of the associated genetic markers. The effect of the markers are estimated and validated so it is possible to predict the genetic values of the candidates of selection based on their genotypes. High density genotyping is expensive so it is usual to genotype the reference population with higher density genotyping chips and to genotype the candidates of selection with lower densities genotyping chips. Genotype imputation is then applied to expand the genotyping data of the candidates, improving the selection intensity and reducing the costs. In this work three imputation softwares, Beagle v4.1, Minimac v3 and Fimpute v2.2, were used to impute genotypes from a lower to a higher density chip using genotyping data from 233 sires of Hereford and Bradford bovine breeds from the south region of Brazil. High-density genotyping data (777k markers) were available for all samples so lower density data (50k markers) could be obtained and the accuracies of the softwares could be measured. Results show that the softwares were able to impute above 94% of all imputable markers. The correctness of the imputation varied from 86% to 94%. The performance varied from 26.9 to 378.1 markers per second, using a sample of the data from chromosome 1. Overall, all three softwares showed good performance and appear to be good choices for the imputation of genotypes to use in GS.

Detection of potential genetic variants affecting gene function in Guzerat cattle

Adhemar Zerlotini^{1¶}, Nedenia Bonvino Stafuzza^{2¶}, Francisco Pereira Lobo¹, Michel Eduardo Beleza Yamagishi¹, Tatiane Cristina Seleguim Chud², Alexandre Rodrigues Caetano³, Danísio Prado Munari², Dorian J. Garrick⁴, Marco Antonio Machado⁵, Marta Fonseca Martins⁵, Maria Raquel Carvalho⁶, Marcos Vinicius Gualberto Barbosa da Silva⁵

¹Embrapa Informática Agropecuária, Campinas, São Paulo, Brazil, ²Departamento de Ciências Exatas, Universidade Estadual Paulista, Faculdade de Ciências Agrárias e Veterinárias, Jaboticabal, São Paulo, Brazil, ³Embrapa Recursos Genéticos e Biotecnologia, Brasília, Distrito Federal, Brazil, ⁴Department of Animal Science, Iowa State University, Ames, Iowa, United States of America, ⁵Embrapa Gado de Leite, Juiz de Fora, Minas Gerais, Brazil, and ⁶Universidade Federal de Minas Gerais, Minas Gerais, Brazil

Guzerat is a dual-purpose breed recognized for important traits to its adaptation to adverse tropical environments such as resistance to parasites, heat tolerance and ability to intake forage with low nutritional value. Once genetic variation responsible for this traits has so far not been well characterized, the aim of this study was to identify single nucleotide variants (SNVs) and insertion/deletions (Indels) in Guzerat cattle breed from whole genome re-sequencing in order to characterize loss-of-function variants which could be associated with complex traits in this cattle breed. The genetic variants identified using HiSeq 2000 sequencing platform were classified using the Ensembl Variant Effect Predictor (VEP) tool, based on a subjective classification of the severity of the variant consequence in transcripts defined by the Sequence Ontology (high, moderate, low or modifier). The DAVID v6.7 tool was used for functional enrichment analysis using the list of genes predicted to be function or activity potentially affected by genetic variation (high and moderate effect) against the background set of bovine genes. The Gene Ontology (GO) enrichment analysis were performed considering 5% FDR threshold for significance. The KEGG pathway database was used to investigate the enriched metabolic pathways ($P<0.01$). Approximately 16.9 million genetic variants (14,588,361 SNVs and 2,322,894 indels) were identified, of which 45,452 (43,922 SNVs and 1,530 indels) were predicted to have high and moderate effect on gene function. The data set comprised 9,922 Ensembl ID because many genes showed a high number of genetic variants. A total of nine GO terms ($FDR<0.05$) were over represented with genetic variants, with the following molecular functions: olfactory receptor, GTPase regulator, nucleoside-triphosphatase regulator, Ras guanyl-nucleotide exchange factor and Rho guanyl-nucleotide exchange factor activity. These enriched terms may reflect the selection history of this Guzerat population and could be associated with some phenotypic traits. The KEGG pathways ($P<0.01$) that could be affected by genetic variants were fatty acid metabolism, ECM-receptor interaction, complement and coagulation cascades, olfactory transduction, aminoacyl-tRNA biosynthesis and lysosome pathways. Because only three animals were sequenced, the over-represented GO terms and KEGG pathways should be carefully interpreted. However, these results provide important genomic information to investigate the genetic mechanisms underlying traits of interest in Guzerat cattle and to improve genomics-based breeding tools.

Characterization of the probiotic and stress resistance-related genes of *Lactococcus lactis* subsp. *lactis* NCDO 2118 through comparative genomics and *in vitro* assays

Oliveira, LC¹; Saraiva, TDL¹; Figueiredo, HCP²; Pereira, UP³; Silva, BC¹; Silva, WM^{1,5}; Silva, A⁴; Azevedo, V¹; Soares, SC^{1,6}

¹Institute of Biologic Sciences, Federal University of Minas Gerais, Belo Horizonte - MG, ²Oceania Laboratory of Fisheries Ministry-Federal University of Minas Gerais, Belo Horizonte, MG, ³State University of Londrina - UEL, Londrina - PR, ⁴Institute of Biological Sciences - UFPA, Belém - PA, ⁵National Institute of Agribusiness Technology, INTA, Argentina, ⁶Federal University of Triângulo Mineiro - UFTM, Uberaba - MG

Lactococcus lactis is highly important for its use in the production of many fermented products and macromolecules and its application in health improvement once it is a Generally Regarded As Safe (GRAS) species. In this scenario, *Lactococcus lactis* subsp. *lactis* NCDO 2118 (herein, NCDO2118) stands out as a xylose fermenter and GABA (gamma-aminobutyric acid) producer strain isolated from frozen peas. However, despite of those important characteristics, few is known about the mechanisms involved in probiotic effects of this strain. In this work, we have sequenced and manually curated the genome of NCDO 2118. Besides, we have compared the genome of NCDO 2118 with those of 7 additional *L. lactis* subsp. *lactis*, 6 *Lactococcus lactis* subsp. *cremoris* and 2 *Lactococcus garvieae* strains. We used the software Gegenees, Mauve and BRIG to perform phylogenomics, gene synteny and circular genome comparisons between the above mentioned species. Additionally, we have used GIPSY and Phast to predict genomic islands (GEIs) and phages, respectively. In phylogenomics analyses, a high similarity was observed between NCDO 2118 and *L. lactis* subsp. *lactis* KF147 (herein, KF147), both isolated from plants. Besides, the gene synteny analyses have clearly shown a highly conserved gene order between both strains. We have also predicted 9 genomic (GEI), 5 metabolic (MI), 4 symbiotic (SI) and 3 miscellaneous islands (MSI - region harbouring metabolic and symbiotic factors). The most prominent genomic island is MSI 2, which presents the biggest region of deletion in all *Lactococcus*, except for NCDO 2118 and KF147. We have also identified 19 phage regions, 3 bacteriocins (from classes I, II and III), 25 acidic and 14 bile stress resistance genes. Finally, although the strain is resistant to Vancomycin, Oxacillin and Amikacin *in vitro*, no antibiotic resistance related gene was identified in putatively horizontally acquired regions. Altogether, the high degree of similarity between all strains point that the SIs commonly shared by both NCDO 2118 and KF147 were responsible for the close relationship in phylogenomic analyses and probably for the adaptation of those strains to plants. The MIs, on the other hand, are highly conserved between all strains, which is an expected feature given the use of those strains in metabolic processes in industry. Finally, the three classes of bacteriocins may have an important role against invasion of competing strains or influencing the host immune system, which may be involved in the probiotic characteristic of this strain.

Financial Support: CAPES, CNPq e FAPEMIG.

A comparison of two pipelines for metagenomic 16S rRNA using Ion Torrent (PGM) Sequencing Platform

Suzana Eiko Sato Guima, Daniel Vasconcelos Rissi, Rodrigo Matheus Pereira
FCBA - UFGD (*Universidade Federal da Grande Dourados*), UPPR (*Universidade Federal do Paraná*) and FCBA - UFGD

Targeted metagenomics is a powerful tool for inferring the phylogenetic distribution of microbial community in samples from different sites or time series. Several pipelines for metagenomic analysis have been developed, and they present different sets of softwares, which makes it difficult to choose the most appropriate one for metagenomic analysis. Here, the goal is to present a comparison between BMP (Brazilian Microbiome Project) and MICCA (Microbial Community Analysis) pipelines for metagenomic 16S rRNA of Ion Torrent Sequencing Data. Microcosm experiments were performed using the native forest soil collected from the experimental farm of EMBRAPA Agropecuária Oeste with weekly applications of fipronil (final concentration of 200µg/kg of soil) for four weeks. Metagenomic library was prepared using V4 and V5 hypervariable regions of the 16S rRNA gene from the extracted microcosm DNA. Sequencing was carried on Ion Torrent PGM Hi-Q using Ion 318 V2 chip. Primers and adapters of the raw reads were removed using FastX-Toolkit software v.0.0.14. Reads were filtered using Sickle software v.1.33, discarding reads shorter than 150bp and trimming them if its quality drops below Q23. From the filtering step, both pipelines were run separately. Alpha diversity and rarefaction curve were processed using QIIME scripts. BIOM format files from both pipelines were statistically analyzed on STAMP software. From the total (321,561) of raw reads, 94,416 sequences were filtered. Rarefaction curve for BMP achieved the plateau while MICCA did not. This result is related to the number of OTUs (Operational Taxonomic Units) clustered in each pipeline. BMP has a lower number of OTUs because it discards singletons (OTU represented by a unique DNA sequence) while MICCA does not, resulting in a higher number of OTUs. The number of classified genus in MICCA was higher than in BMP. Although the results at the genus level were different, results between both pipelines were statistically similar at the family and order level. Therefore, both pipelines can infer a similar bacterial taxonomic diversity from phylum to family level. It is suggested the use of BMP for the bacterial diversity analysis considering the most common genus, and the use of MICCA for bacterial community analysis considering rare genus and species.

Insights into *Klebsiella pneumoniae* type VI secretion system regulation

Victor Barbosa, Leticia MS Lery

Laboratório de Microbiologia Celular – IOC/Fiocruz

Klebsiella pneumoniae is a Gram-negative bacterium responsible for many acute infections, mainly in the urinary and respiratory tracts. These infections represent a big challenge for public health as there are multiple antibiotic resistant strains circulating around the world, including in Brazil. Thus, the understanding of *K. pneumoniae* virulence mechanisms is still required and may, in the future, lead to novel approaches for interfering at the infection process, such as the development of more efficient and specific drugs. Recently, it was suggested that the type VI secretion system (T6SS) of *K. pneumoniae* is important to its pathogenicity. Indeed, T6SS provides competitive and adaptative advantages for bacteria that possess it. As T6SS is present in a wide variety of proteobacteria, it is not surprising that the T6SS shows different regulatory mechanisms, such as quorum sensing, biofilm formation, iron limitation, oxidative stress, changes in osmolarity and temperature as well as the independent regulation for the expression structural components and of the T6SS effectors. Up to now at least 24 transcriptional regulators of T6SS are known in species of *Vibrio*, *Pseudomonas*, *Burkholderia* and *Escherichia* genus. However, there are no studies concerning T6SS regulation in *K. pneumoniae*. Our group has previously observed that *K. pneumoniae* strain Kp52.145 expresses T6SS genes *in vivo*, however the signal triggering such mechanism has not been identified. In this work we aimed to get insights into such regulation. In order to achieve this aim, we used a computational approach to identify possible transcriptional regulators of T6SS in *K. pneumoniae*. In order to infer promoter regions, we predicted potential transcription start sites in the three T6SS loci encoded in the genome of this strain, using BPROM algorithm. Then, we further analyzed the 13 promoter regions predicted in order to identify putative transcriptional regulators binding sites. For that, the -500 bp sequences were analyzed with Virtual Footprint software against Prodoric database consisting of 59 protein binding site patterns. In overall, 523 putative binding sites of 27 different regulators were predicted. After manual curation, eight regulators and their binding sites are highlighted. They are: Fnr | *Escherichia coli*, Fur(8mer) | *Escherichia coli*, Fur | *Pseudomonas aeruginosa*, OmpR | *Escherichia coli*, OxyR | *Escherichia coli*, OxyR (SELEX) | *Escherichia coli*, RcsAB | *Escherichia coli*. Most of those regulators are part of two-component systems regulated by known environmental stimuli. We are currently performing validation experiments in order to verify computational predictions.

AMP-Identifier: A Unix shell script for antimicrobial peptide identification

Bezerra-Neto, João Pacifico^{1,2}; Santos, Mauro Guida²; Benko-Iseppon, Ana Maria¹

¹*Laboratory of Plant Genetics and Biotechnology, Genetics Department, Universidade Federal de Pernambuco, Av. Prof. Moraes Rego, 1235, 50.670-423, Recife, PE, Brazil;*

²*Laboratory of Plant Physiology, Department of Botany, Universidade Federal de Pernambuco, Av. Prof. Moraes Rego, 1235, 50.670-423, Recife, PE, Brazil*

With the advent of experimental high-performance platforms, in particular next-generation sequencing (NGS) optimized assay systems, advanced bioinformatics approaches have enabled comprehensive and maximized studies of eukaryotic genomes in a quick and economically viable manner. In plants, for example, transcriptome studies have been used for quantitative analysis of thousands of expressed genes related to germination, growth and development, flowering, and conditions of biotic and abiotic stresses, allowing the understanding plant response mechanisms against stresses. The identification of gene families in the huge amount of new data has been facilitated by bioinformatic methods and by the availability of several online repositories. The main computational methods developed for identifying AMPs (Antimicrobial Peptides) on a genome-wide scale involved in silico approaches to evaluate their amino acid composition and structure. Scripts, written in Unix shell or other scripting languages such as Perl, can be seen as the most basic form of pipeline framework. The AMPs generally present between 12 and 50 amino acids, including the presence of disulfide bond and/or cyclization of the peptide chain. These peptides have a variety of antimicrobial activities ranging from membrane permeabilization to action on a range of cytoplasmic targets. To improve the identification of AMPs at omic level, we developed a Unix shell script to integrate other analysis tools to find AMPs from HMM models. This automated pipeline adopts classification based on HMM models cataloged in CAMP (www.camp3.bicnirrh.res.in) database, search based on HMMER tool in some cases translation from nucleotide to amino acids for genomic data input using TransDecoder tool. This script gathers all tools and HMM database, building all commands to execute the analysis, asking the user just about input and parameters of AMP search. To evaluate the script efficiency we downloaded the *Arabidopsis thaliana* genome to run as input, using a cutoff of 0.00001. We obtained 32 AMP families identified on *A. thaliana* genome under the selected cutoff, using 89 HMM models. For predicted Defensin sequences we found that against GenBank, only 50% were confirmed by BLAST and CD-search tools, indicating that our tools may identify sequences not classified as AMPs in conventional alignment approaches. The here presented tool facilitates AMP global identification and its usability, once Unix environment may be a challenge for most biologists, since the implementations are based on Linux command lines, often requiring some knowledge.

Financial support: CNPq, CAPES.

Genome mining of biosynthetic gene clusters in *Nostoc* sp. CACIAM 19, a cyanobacterium from an Amazonian environment

David Batista Maués¹, Alex Ranieri Jerônimo Lima¹, Pablo Henrique Gonçalves Moraes², Andrei Santos Siqueira¹, Leonardo Teixeira Dall’Agnol³, Evonnildo Costa Gonçalves¹

¹Laboratório de Tecnologia Biomolecular, Instituto de Ciências Biológicas, Universidade Federal do Pará

²Laboratório de Farmacognosia 2, Campus São Luís, Universidade Federal do Maranhão

³Campus Bacabal, Universidade Federal do Maranhão

Cyanobacteria comprise the largest, most diverse, and the most widely distributed group of photosynthetic prokaryotes, being found in all types of terrestrial and aquatic ecosystems, and constitutes a rich source of compounds of great significance in biotechnology. The *Nostoc* genus, constituted by filamentous heterocystic cyanobacteria, nitrogen fixing, has been utilized as a source of vitamins, proteins and fatty acids, also of secondary metabolites with anticancer, antimicrobial and antiviral activities. The number of these organisms' genomes publicly available has rapidly grown in the last few years. In this sense, genome mining of their biosynthetic gene clusters has become a key approach for novel compound discovery. In order to evaluate its biotechnological potential, we obtained the draft genome of *Nostoc* sp. CACIAM 19. This strain was isolated from a water sample from Bolonha lake, localized in Belém, Pará. After DNA extraction of the cyanobacterial non-axenic culture, two sequencing runs were performed on the GS FLX 454 (Roche Life Sciences) platform using non-paired libraries, and one sequencing run was carried out on the Illumina MiSeq platform using a paired-end library with 150 bp read length. A co-assembly of all reads was performed by Newbler 2.9 and the resulting scaffolds were binned with MaxBin 2.2.1 and then submitted to the antiSMASH 3 tool. The predicted clusters were manually evaluated using the data banks Pfam, Uniprot, NCBI and Interproscan. Thirty-two clusters of biosynthetic genes for secondary metabolites were identified, among them: 12 exopolysaccharides biosynthesis clusters, 3 bacteriocins biosynthetic clusters, 1 cyanobactin biosynthetic cluster, 4 fatty acid production clusters, 4 hydrocarbon production clusters, 1 microviridin producing cluster, 1 phosphonate biosynthetic cluster and 6 NRPS/KS clusters. The terpenoid and hydrocarbon clusters, whose products can be utilized in the chemical industry, showed up to 35% of similarity with the clusters present in *Nostoc punctiforme* PCC 73102. Such results show the potentiality and diversity of secondary metabolites produced by the amazon cyanobacteria *Nostoc* sp. CACIAM19, which can be explored in diverse sectors of the biotechnology industry.

Support: CNPq, CAPES, CIT-IEC, FAPESPA.

Ab initio characterization of promoter regions based on Conditional Random Fields

Ígor Bonadio, Mauro de Medeiros, Alan Mitchell Durham

Programa de Pós Graduação em Ciência da Computação da USP, Programa de Pós Graduação em Bioinformática da USP, Instituto de Matemática e Estatística da USP

Gene prediction aims to find the location of genes in a genome. However, current gene prediction programs just identify the coding regions and not the promoter regions. Identifying the promoter region involves correctly locating the transcription start site (TSS), which is a difficult task due the lack of a strong signal around this site. Many techniques were developed but their number of false positives is too high for practical use. In this project we propose a new method based on Conditional Random Fields (CRF) that presents a much better prediction rates than previous algorithms. With our approach we are able to predict not only the coding region but also to approximately locate the TSS, TATA-box and CCAAT-box. The use of CRFs enables us to effectively combine the annotation generated from a traditional GHMM-based gene predictor with information about the nucleotide composition of the intergenic region, the distance distribution between start codons and TSSs and between TSSs and TATA-boxes. We validated our methodology using the PlantProm database, which have annotation of the promoter region of 579 plant genes (monocots and dicots) including experimentally verified TSSs, putative TATA-boxes and putative CCAAT-boxes. Our approach was able to approximate the TSS location with a much higher precision than other approaches. In particular, 74.95% of the TSSs were identified with maximum error of up to 30 nucleotides, 58.03% with an error of up to 20 nucleotides, and 35.92% with a maximum error of only 10 nucleotides. This first modeling of the promoter region can help reduce false positives in the process of *ab initio* TFBS discovery. We plan in the near future to investigate more sophisticated models of promoter regions with other signals such as Y-Patch, DPE, MTE, INR, DCE and MDE.

Study of Chromatin Remodeling in Colorectal Cancer Progression

Simone Nantes de Aquino, Nicole M. Scherer, Mariana Boroni

Laboratório de Bioinformática e Biologia Computacional, Instituto Nacional de Câncer

Colorectal cancer (CRC) it is the result of an accumulation of genetic and epigenetic changes in colon epithelial cells, which converts them into adenocarcinomas. Many of these tumors start from polyps, benign lesions that may occur on the inner wall of the large intestine and if untreated, can develop to an invasive cancer. Epigenetic changes play an important role in cellular differentiation process, allowing cells to be phenotypically diverse despite containing the same genetic content. Selective modifications of histones have been shown to act together with DNA methylation, both resulting in the modification of the chromatin conformation, which therefore influences the expression of genes. These alterations influence the nucleosome positioning along the DNA, and it is a crucial factor of chromatin accessibility. Like gene mutations, these alterations can also contribute to the pathogenesis and molecular heterogeneity of tumors. Thus, increased understanding of the gene expression regulation from epigenetic context during the CRC progression may contribute to the development of new epigenetic markers that can be applied to diagnostic, tissue invasion tendency and metastasis, prognosis or response to chemotherapy agents. In order to study the chromatin structure of CRC genomes, we have downloaded data from the public database SRA of six paired samples, normal and tumor, of the study SRP065259: SRR2810481 to SRR2810486. This data were generated by using MNase-Seq followed by enrichment of transcription start sites (TSS) regions. This approach leads to the mapping of nucleosome positions in TSS regions by treating chromatin with micrococcal nuclease (MNase), which preferentially digests linker DNA, followed by paired-end sequencing of undigested DNA fragments (MNase-seq) that came from TSS regions. After the download, the quality of the reads were analyzed using FastQC tool. Low quality reads were removed using Trimmomatic algorithm. Then, reads were aligned to the human genome (GRCh37 version) with BWA program. The RSeQC package was used to evaluate the quality of alignments and only reads with single alignment were accepted for subsequent analysis. The positioning of nucleosomes was determined using MACs2 tool, generating the following numbers of peaks per sample: 51804 (SRR2810481); 53275 (SRR2810482); 52550 (SRR2810483); 58018 (SRR2810484); 55919 (SRR2810485) and 57404 (SRR2810486). All changes found in primary metabolic pathways and related tumor signaling will be further characterized using the tools KEGG, GO and Reactome, in order to determine the impact of chromatin remodeling in the progression of the CRC.

Financial support: Ministério da Saúde

Taxonomic identification of metagenomics reads based on sequence features by Support Vector Machines (SVM)

Tahila Andrigatti¹, Ney Lemke¹

¹*Universidade Estadual Paulista Júlio de Mesquita Filho, Departamento de Física e Biofísica, Botucatu, São Paulo, Brasil*

The acknowledgement of the importance of microbiota composition is increasing steadily after the advent of metagenomics. This approach allows sequencing and analyzing genetic material from a microbial community without the need of microbial culture. Since 99% of microorganisms are not culturable, metagenomics is the standard methodology to investigate microbiomes composition and dynamics. However, the actual output data of metagenome sequencing consists of a bunch of DNA fragments originated from various microorganisms. Moreover, the lack of reference genomes in databases challenges taxonomic identification of unknown organisms in these samples and unbiased estimates for the performance of the proposed methodologies. In this work, we evaluated the predictive power of Support Vector Machine (SVM) learning tool on taxonomic classification in phyla of unknown metagenomics DNA reads. To simulate the identification of unknown microorganisms, we used Gammaproteobacteria sequences excluding *Escherichia coli* as the training set in SVM. From the trained model, we classified the sequences of *E. coli* and analyzed if they were correctly assigned on Gammaproteobacteria group. The tests were performed for 100, 400 and 1000 bp test sequences to evaluate the influence of size on the prediction. The simulations were performed using the following DNA measurements as SVM input: GC content, di, tri and tetraplet entropy, di, tri and tetranucleotides frequencies (2, 3 and 4-mers), dinucleotide abundance and tetranucleotide derived z-score correlations (TETRA). We tested sets of measurements composed by all parameters but excluding one to compare the relative impact of each measure. We found that the groups which excluded TETRA are less suitable for the most of sizes tested, specially for 100 bp. The other groups showed AUC values higher than 0.7 for prediction of unknown sequences. The use of sequence features is an interesting approach to characterize sequences of not fully sequenced organisms.

Cell cycle and metabolism related candidate human synthetic lethal network

Sandeep Tiwari¹, Thiago Luiz de Paula Castro¹, Núbia Sei eert¹, Debmalya Barh²,
Vasco AC Azevedo¹

¹PG program in Bioinformatics (LGCM), Institute of Biological Sciences, Federal University of Minas Gerais, Belo Horizonte, MG, Brazil. ²Centre for Genomics and Applied Gene Technology, Institute of Integrative Omics and Applied Biotechnology, Nonakuri, Purba Medinipur, West Bengal-721172, India

Synthetic lethal strategy can be used as a power full method in the development of anticancer drug for the fighting against cancer. It is evident from various published literatures that identification of synthetic lethal gene pairs and targeting one of them can be an effective approach in developing targeted anti-cancer drug where the cancer cells carry a mutation in the other pair of the synthetic lethal duo. The strategy is successfully applied in BRCA1 mutant breast cancer by targeting PARP due to the synthetic lethal relationship between these two genes. Identification of synthetic lethal pairs requires genome wide screening and therefore in human is a difficult task. For yeast, nematode, and fly synthetic lethal screening have been reported using in vitro mutagenesis. However, in human, such experimentally validated reports are limited. Although comparative genomics and phylogenetic relationship based predicted synthetic lethal relationship are available. Here, novel bioinformatics approaches were used, based on our strategy, David, ToppGene, and Osperry analysis, we found cell cycle related genes MDM2:MDM4 that are reported to be synthetic lethal in Human, are binding to each other. We have identified cell cycle and metabolism related candidate human synthetic lethal network using yeast proteome. Our analysis is also verified using reported synthetic lethal interactions in human and shows the applied method is enough powerful to screen new synthetic lethal relationships in human. Many key nodes of our identified network are involved in cancer. Therefore, this new synthetic lethal network should be further explored for development of anticancer strategy. These candidate synthetic lethal genes required future experimental validation. Most of the identified candidate synthetic lethal genes are involved in tumorigenesis process and therefore selective targeting of a partner of a pair may provide effective anti-cancer therapy.

The discovery of novel multiple small deletions within human coding genes associated to known lung cancer pathways

Gabriel Wajnberg^{1,2}, Raphael Tavares da Silva¹, Nicole de Miranda Scherer³, Carlos Gil Ferreira⁴ e Fabio Passetti^{1,2}

1. Laboratory of Functional Genomics and Bioinformatics, Oswaldo Cruz Institute, FIOCRUZ, Rio de Janeiro, Brazil., 2. Systems and Computational Biology Graduate Program, Oswaldo Cruz Institute, FIOCRUZ, Rio de Janeiro, Brazil., 3. Laboratório de Bioinformática e Biologia Computacional, Instituto Nacional de Câncer (INCA), Rio de Janeiro, Brazil., 4. Clinical Research Coordination, Instituto Nacional de Câncer (INCA), Rio de Janeiro, Brazil.

Deletions occur naturally in the coding genes of healthy human populations, which, in turn, may affect the protein product sequence. Although the identification of deletions is usually performed using DNA data, the use of transcriptome data is a promising strategy. In addition, the 1000 genomes project can be used as source of data to search for deletions in different healthy human populations. We developed an innovative method to identify small deletions using human transcriptome data. Here, we present the detection and analysis of small deletions in 22 matched tumor and normal lung cancer RNA-Seq data using the human genome GRCh37/hg19 as reference. Using this strategy, we identified 1,778 small deletions using the assembled transcripts in these lung cancer samples. We confirmed 928 (52%) of these on a selected list 66 genome sequences comprising 3 from each of the 22 populations available in the 1000 genomes project. We identified 398 (22%) from the 928 set of small deletions predicted to change the known reading frame in 194 different genes. If considered deletions detected in both matched normal and tumoral samples, 53 (13%) were identified. We identified three altered pathways with significant posterior probability to be overrepresented performing a Gene Set Enrichment Analysis (GSEA): positive regulation of cell proliferation, positive regulation of Notch signaling pathway, and negative regulation of apoptotic process. These findings are supported by other studies. For example, the positive regulation of cell proliferation path (GO:0008284) has been already been previously identified as overrepresented in using lung cancer samples. The *TNFSF13B* is described to influence this pathway and it has been already identified with high expression in lung cancer. We identified a 88 nucleotide small deletion in the *TNFSF13B* in 8 tumor samples from different patients, providing a strong support of our findings. According to our analysis, this deletion affects the main TNF domain encoded by the canonical protein. In conclusion, here we describe the identification of a group of small deletions, which may contribute the reduced control of previously described pathways associated to lung cancer, improving the knowledge of the lung cancer biology. Financial support: INCA/MS, FIOCRUZ, CAPES, Fundação do Câncer, FAPERJ and CNPq.

Identification of somatic mutations in prostate adenocarcinoma with Gleason score 7 and 8 and their associations with biochemical recurrence

Isabella T J e Meira¹, Bruna D F Barros¹, Rodrigo F Ramalho¹, José E Kroll², Renan Valieres¹, Sandro J de Souza², Isabela W da Cunha³, Gustavo C Guimarães⁴, Dirce M Carraro¹, Elisa N Ferreira¹

International Research Center (CIPE) - A.C. Camargo Cancer Center¹, Brain Institute – Federal University of Rio Grande do Norte², Department of Anatomic Pathology - A.C. Camargo Cancer Center³ and Department of Urology - A.C. Camargo Cancer Center⁴

Prostate cancer is a heterogeneous and multifocal disease. In general, it presents an indolent behavior, being asymptomatic in many cases. However, in some cases, the tumor rapidly progress to metastatic disease leading to patient death. The Gleason score (GS) is the main prognostic factor for localized disease and its classification is based on the sum of the primary and the secondary histological pattern, ranging from 2 to 10. Despite being the main classification method, GS alone does not provide accurate information about patient outcome, since patients with the same score can have different behaviors, particularly in intermediate GS (7-8). Therefore, in this study, we investigated genomic alterations in patients with intermediate GS that presented biochemical recurrence (BCR) compared to patients without recurrence, aiming to improve prognosis and also to reveal biological pathways involved with tumor aggressiveness. Thirty-two prostate adenocarcinoma patients with intermediate GS (7-8) that underwent radical prostatectomy at the A.C. Camargo Cancer Center, were selected and divided into 2 groups, 15 patients with and 17 patients without BCR, respectively. Using the TruSeq Custom Amplicon kit and the NextSeq platform (Illumina), we performed targeted-sequencing of 58 carefully selected genes, including 23 genes frequently mutated in prostate cancer and 35 genes mutated in other solid tumors, in 64 paired tumor/normal samples, generating $\geq 1000X$ average coverage per sample. Data were analyzed by the TruSeq Amplicon tool, which uses an aligner based on the Smith-Waterman algorithm and the “Somatic” variant caller. To identify somatic mutations, we compared tumor/normal samples using Varseq software and selected SNVs and Indels with a minimum coverage of 100X and a minimum allele frequency of 2% in the tumor, that leads to missense, nonsense, splice site or frameshift alterations. A total of 246 variants were identified, 161 SNVs and 85 Indels, with an average of 8 alterations per tumor (ranging from 0 to 41). Of the 58 genes investigated, 49 were affected in at least one patient (84.5%), suggesting that the custom panel is enriched in genes mutated in prostate cancer. Further analyzes are being conducted to evaluate the number and type of variants in the group of patients with and without BCR and to investigate which genes were preferentially affected and their related biological pathways. Therefore, we expect to obtain a profile of the genes most frequently mutated in prostate cancer and investigate possible associations with clinicopathological characteristics, to improve prognosis and better estimate the risk of recurrence.

Financial support CNPq (459113/2014-3) and CAPES.

Detection and correction mis-assemblies in genome of *Corynebacterium pseudotuberculosis*

Thiago de Jesus Sousa¹, Doglas Parise¹, Diego César Batista Mariano², Daniela Arruda Costa¹, Felipe Luiz Pereira³, Henrique César Pereira Figueiredo³, Artur Silva⁴, Rommel Thiago Jucá Ramos⁴, Vasco Azevedo¹.

¹Laboratory of Cellular and Molecular Genetics, Department of General Biology, Institute of Biological Sciences, Federal University of Minas Gerais;

² Laboratory of Bioinformatics and Systems, Department of Computer Science, Federal University of Minas Gerais;

³ National Reference Laboratory for Aquatic Animal Diseases of Ministry of Fisheries and Aquaculture, Federal University of Minas Gerais;

⁴Institute of Biological Sciences, Federal University of Pará.

Corynebacterium pseudotuberculosis is a bacterium that belongs to CMNR group, which includes the genus *Corynebacterium*, *Mycobacterium*, *Nocardia* and *Rhodococcus*. This species is the etiologic agent of Caseous Lymphadenitis (CLA) in sheep and goats (*C. pseudotuberculosis* biovar *ovis*); and Ulcerative lymphangitis in horses, cattles, buffaloes and camels (*C. pseudotuberculosis* biovar *equi*). The first genome sequenced of this organism was *C. pseudotuberculosis* strain 1002, in 2006, which was deposited in the National Center for Biotechnology Information (NCBI) in 2009. Currently, 43 strains have their genomes available in the NCBI database, in which previous studies have shown that these genomes may contain assembly errors. These findings have been allowed by the evolution of next generation sequencing platforms, which provides high precision and reduced cost data. In addition, through restriction enzymes, the use of genome optical mapping enabled improvements in data assembly. Moreover, high precision contigs sorting and high accuracy data have allowed the detection of large genomic rearrangements. This work aimed to update *C. pseudotuberculosis* strain I19 (CpI19) and *C. pseudotuberculosis* strain 162 (Cp162) genomes. For this purpose, a new sequencing of these strains were done using a 400 pb fragment library in Ion Torrent PGM™ platform. The new CpI19 and Cp162 sequencing generated respectively 376,308,624 bp with coverage 160.98-fold and 473,348,503 pb and coverage 200.09-fold. The scaffolding process was performed using MapSolver software to contigs sorting, which utilizes a strategy based on restriction maps constructed upon known restriction sites recognized by KpnI enzyme *in vitro* by OPGEN, Inc. (Gaithersburg, USA). After the assembling process, a genomic inversion of 1.22MB in CpI19 and 0.85MB in Cp162 were identified. These results showed also a reduction of 146 bp in CpI19 and an addition of 72.215 pb in Cp162. A combined analysis utilizing optical mapping and sequencing data enabled the detection of assembly errors, and genomic inversions as well as genome size inconsistencies in the two previously deposited genomes, showing the optical mapping efficiency

Genomic analysis of opportunistic bacteria from *Herbaspirillum* genus isolated from immunocompromised patients

¹Willian Klassen de Oliveira, ²Michelle Zibeti Tadra-Sfeir, ²Rodrigo Luis Cardoso,
²Emanuel Maltempi de Souza, ²Fábio de Oliveira Pedrosa, ^{1,3}Helisson Faoro

¹Laboratory of Bioinformatics, Professional and Technological Education Sector,
Universidade Federal do Paraná; ²Department of Biochemistry and Molecular Biology,
Universidade Federal do Paraná; ³Laboratory of Gene Expression Regulation, Carlos
Chagas Institute, Fiocruz-PR

Herbaspirillum is a genus of the beta class of the Proteobacteria phylum. Some species of this genus are of biotechnological interest as growth promoters of crops like maize, rice and sugarcane. They are capable to establish endophytic association with these plants and to convert the atmospheric nitrogen (N_2) in a form that could be used by the plants (NH_4), besides to secrete phytohormones. However, strains of *Herbaspirillum* spp. have been isolated from immunocompromised patients, sometimes leading the patient to develop a bacteremia and death. To understand how environmental organisms evolve into a clinical variant and the molecular mechanisms involved in this process we sequenced, assembled and annotated the genome of two clinical strains of the *Herbaspirillum* genus: *Herbaspirillum frisingense* AU14559 and *Herbaspirillum* lineage 2 AU13964. Both strains were isolated from sputum of patients with cystic fibrosis and classified as *Herbaspirillum* based on the 16S rRNA gene sequence comparison. The *Herbaspirillum* lineage 2 was assembled in 10 contigs, with total size of 5.35 Mb, the largest contig has 1.87 Mb, GC % 63.21, N50 of 539.6 kb, L50 of 3 contigs, 3 rRNA operons, 4,880 annotated genes and 65 predicted tRNAs. The Average Nucleotide Identity (ANI) comparison of *Herbaspirillum* lineage 2 showed the higher level of identity to *Herbaspirillum seropedicae* SmR1 (97.59%), indicating that this strain can be included within the *Herbaspirillum seropedicae* species. The *H. frisingense* AU14559 was assembled in 15 contigs, with a total size of 5.44 Mb, the largest contig has 2.38 Mb, GC% 63.11, N50 557.3 kb, L50 of 2 contigs, 3 rRNAs operons, 5,213 predicted genes and 56 predicted tRNAs. The ANI comparison to the genome of the environmental strain, *Herbaspirillum frisingense* GSF30, was 97.39%. Searches made through the BLAST algorithm revealed the absence of the *nifHDK* genes that codes the structural proteins of the nitrogenase complex, indicating that these two organisms are incapable to fix nitrogen. Additional genome comparison studies will provide insights about the evolution of these strains from the environmental to clinical lifestyle.

Supported by: INCT-Fixação Biológica de Nitrogênio, MCTI-CNPq.

Investigation of mutations in the *HBB* gene using the 1000 GENOMES databank

Tânia Carlice-dos-Reis, Jaime Viana, Fabiano Moreira Cordeiro, Greice de Lemos Cardoso, João Guerreiro, Sidney Santos, Ândrea Ribeiro-dos-Santos

Laboratory of Human and Medical Genetics, Institute of Biological Sciences, Federal University of Pará, Belém, PA, 66.075-110, Brazil; Federal Rural University of the Amazon, Capanema Campus, PA, 66.077-830, Brazil; Research Center of Oncology, Federal University of Pará, Belém, PA, 66.073-005, Brazil.

Sickle-cell disease is one of the most common monogenic diseases worldwide, caused by mutations in the HBB gene (β -globin). Due to its high prevalence, various strategies have been developed to better understand its molecular mechanisms. In silico analysis has been increasingly used to investigate genotype-phenotype relationship of many diseases, and the sequences deposited in the 1,000 Genomes database, of healthy individuals, appears to be an excellent approach for this analysis. This study aims to analyze the variations of the HBB gene in the 1,000 Genomes database, as well as investigate the pattern of pathogenicity, and describe the mutations frequencies in the different population groups. The computational tool SnpEff was used to select HBB mutation identified among 2,504 samples from 1,000 genomes. Nucleotide mutations, amino acid changes, allelic and population frequencies, and type of mutations were visualized using the IGV software. The pathogenicity of each amino acid change was investigated using the databases CLINVAR, dbSNP and five different predictors (POLYPHEN, SIFT, PROVEAN, PANTHER e MUTPRED). Pathogenic mutations of HBB, according to the predictors, that were not identified on CLINVAR database were 3D modeling in the PDB database, to infer the effect of the mutation on protein function. Were found 20 different types of mutations in 209 individuals, where 173 subjects had missense mutations. The African population group presented the highest number of mutated individuals 153 individuals, and European presented the least (9 individuals). According to the results, 70% of the mutations were pathogenic. The constructed 3D model allows to visualize residues that have undergone mutation and location of each of the protein. It is concluded that approximately 8.3% of phenotypically healthy individuals from the database 1,000 Genomes have some mutation in the HBB, of which 70% are pathogenic. The mutations are unequally distributed among the populations, being the most affected the African population (73.2% of subjects) and European population the less affected (4.3% of subjects). Pathogenic mutations with greater allele frequencies (rs334, rs33930165 and rs33950507) are known to cause sickle-cell disease and β -thalassemia.

An hierarchical classification system for beta-lactamases

Melise Chaves Silveira¹, Fábio Mota¹, Rodrigo Jardim¹, Rangeline Azevedo daSilva¹, Marcos Paulo Catanho de Souza², Ana Carolina Ramos Guimarães², Antônio Basílio de Miranda¹

Laboratório de Biologia Computacional e Sistemas¹; Laboratório de Genômica Funcional e Bioinformática²

NGS sequencing projects are revealing a growing number of unique and naturally occurring beta-lactamases, enzymes able to irreversibly inactivate beta-lactams antibiotics, the major option to treat bacterial infections. This diversity and the major clinical impact of beta-lactamases led to several attempts to achieve a representative classification system. Ambler's structural classification (1980) is the most used, but it does not represent the evolutionary relationships between these enzymes, the reason Hall and Barlow suggested a hierarchical organization of it (2005). In this work, we propose a system to identify and hierarchically classify beta-lactamases, considering structure and sequence characteristics, prioritizing confidence when attributing function and class for a given protein. Beta-lactamases (EC 3.5.2.6) primary and tertiary structures were downloaded from PDB. Structural and sequence hierarchical clustering tests using MaxCluster and BLASTCLUST programs were performed to achieve Ambler's classification as reviewed by Hall and Barlow. According to it, serine beta-lactamase classes A, C and D were renamed to SA, SC and SD, and the metallo beta-lactamases were divided in MB and ME, and in the next level MB is divided in subclasses B1 and B2. After achieving the five initial clusters, we constructed Hidden Markov Models profiles for each one using the HMMER package and the protocol HMM-ModE. The profiles were tested for class specificity (CE), beta-lactamase function specificity (FE) and beta-lactamase function sensitivity (FS) through in house scripts, using a beta-lactamase dataset constructed by us, the CATH database and the Swiss-Prot database, respectively. Single linkage hierarchical clustering using beta-lactamase structures formed the five clusters expected; sequence clustering under a 50% similarity threshold separated groups B1 from B2; and using a 60% similarity threshold a novel level is suggested, where classes SA, SD, B1 and ME were each divided in two, corroborating previous studies by other groups. All profiles obtained 100% CE. Together, the profiles have 87% FS using the Gene Ontology/Swiss-Prot annotation. We noticed that the remaining 13% is actually comprised by proteins without beta-lactamase activity. MB and ME profiles displayed 100% FE, as either SA, SC and SD after to apply phylogenetic methods to determine new score thresholds and establishing "gray zones" for searches. With these procedures we have achieved a system for the identification and classification of beta-lactamases exclusively on the basis of different levels of protein structures, that also reflects their evolutionary relationships. The system was curated and proved to be efficient, able to be applied in large scale.

Funding support: Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES)

Draft genome sequence of the extremophile endemic marine antarctic yeast *Metchnikowia australis*

Heron Oliveira Hilario¹, Thiago Mafra Batista¹, Rennan Garcia Moreira¹, Valéria Martins Godinho², Carlos Augusto Rosa², Luiz Henrique Rosa², Glória Regina Franco¹

Departamento de Bioquímica & Imunologia, Universidade Federal de Minas Gerais¹,
Departamento of Microbiologia, Universidade Federal de Minas Gerais²

Antarctica, the earth's fifth biggest continent, is an harsh environment. Almost all of its 14.000.000 km² surface is always under ice, originated from the snow falling constantly since past ages. In the winter this ice sheet grows over the Antarctic Circle, till the the northernmost tip of the continent, where situates the Antarctic Peninsula. These geocycles selected diverse survival strategies for life in the region, and only the organisms adapted to deal with extreme situations are capable of flourishing, in special, island microorganisms which are poorly characterized due to their difficult cultivation conditions. As one of the less explored places on earth, Antarctica potentially harbors indigenous microorganisms which may present features that could be exploited by the scientific and medical communities, such as novel candidates for antibiotics, anti-freezing agents and other pharmacological composites. The Mycoantar project integrates the annual expeditions of PROANTAR (Brazilian Antarctic Program), investigating this biodiversity and has already isolated many fungi from the antarctic region. The yeast endemic *Metschnikowia australis* was isolated from marine macroalgae from Antarctica and after diverse microbiological characterization, it was further sequenced for deeper bioinformatic investigation. Paired-end libraries were constructed using Nextera XT DNA Kit, producing fragments with mean of 1,167 bp that were sequenced in the Illumina MiSeq and fragments of 550 bp that were sequenced in the Illumina HiSeq 2500. A total of 1,585,122 reads (2x301) was generated by MiSeq with an estimated genome coverage of 35x, and 103,312,458 reads (2x101) were generated by Hiseq with an estimated genome coverage of 745x. The genome was assembled using SPADES 3.9.1, with default parameters. The estimated genome size is 14,356,710 bp comprising 160 contigs with mean of 89,729 bp, and the longest contig length of 1,116,518 bp, N₅₀ value of 542,232 and GC content of 47.2%. The predictor Maker2 was able to find 4,442 Open Reading Frames. The search for sequence similarity against the non-redundant database from NCBI revealed hits of these ORFs with 4,348 distinct proteins. Comparative analysis of orthologous proteins present in the genome of *M. australis* and other *Metchnikowia* species (*M. bicuspidata* and *M. fructicola*) revealed several shared clusters and 163 singletons, however, six clusters were composed solely of *M. australis* proteins. Downstream analysis will be carried out to deep investigate genes from *M. australis* and their involvement in the cold adaptation.

Diversity analysis of Howler monkey (*Alouatta spp.*) fecal microbiota

R. R. A. Franco^{1,2}, L. F. Martins¹, A. M. Thomaz^{1,2}, J.B.Cruz⁴, J. C. F. de Oliveira⁴,
J. C. Setubal^{1,2} and A. M. da Silva^{1,2}

¹Departamento de Bioquímica, Instituto de Química, USP; ²Programa de Pós-Graduação
Interunidades em Bioinformática, USP; ³Departamento de Ciências Biológicas,
Universidade Federal de São Paulo; ⁴Fundação Parque Zoológico de São Paulo

Howler monkeys (*Alouatta spp.*) are endemic species from the Atlantic Forest biome that can be found in primary and secondary forests and even in small forest fragments. Their diet is based on tree leaves and fruits, depending on the season. This study aims to investigate the diversity of gastrointestinal bacterial community of howler monkeys that inhabit São Paulo Zoo Park, both in captivity and non-captivity, to correlate possible differences between their respective microbiotas and diets. We have collected a total of 25 fecal samples from captive and non-captive individuals at different seasons in 2013-2015. Total DNA extracted from the samples were then analyzed by 16S rRNA gene V3-V4 amplicon sequencing using the MiSeq-Illumina platform. The obtained sequences were used for alpha- and beta-diversity estimates as well as for phylogenetic profiling using mostly the QIIME package. Our initial results point to differences both in the microbial community profile and diversity between the two groups. The phyla Spirochaetes and Elusimicrobia were detected only in captive animals while Tenericutes and Melainabacteria were present only in the microbiota of non-captive individuals. Nevertheless, Bacteroidetes and Firmicutes showed high abundance (~60-70%) in both groups. The microbiota of the non-captive group was richer than the captive one, and presents a large fraction (~70%) of OTUs that were unclassified at the genus level. Among the identified genera, we observed an abundance of *Bacteroides* and *Prevotella* in the microbiota of captive animals. In humans, these two genera have been related to diets high in fat/protein and carbohydrates/fiber, respectively.

Supported by FAPESP, CNPq and CAPES.

Spacial Organization of Genomes: Insights on coordinated regulation of Biological Pathways

Luís Henrique Trentin de Souza¹, José Miguel Ortega¹

UFMG – Universidade Federal de Minas Gerais

Human genome sequencing was one of the biggest breakthrough in biomedical history. However, not all the expected information was found in genome linear sequence. It is been widely accepted that genome structure also plays an important role in its function. So the development of chromosome conformation capture (3C) technology and the subsequent genomic variants thereof have enabled the analysis of nuclear organization at an unprecedented resolution and throughput. The technology relies on the original and, in hindsight, remarkably simple idea that digestion and religation of fixed chromatin in cells, followed by the quantification of ligation junctions, allows for the determination of DNA contact frequencies and insight into chromosome topology. Hi-C is an unbiased (all *vs.* all) chromosome conformation capture technique that generates a list of intra and inter-chromosomal contacts which can help to solve the genomic tridimensional structure. Analysing Hi-C data retrieved from public repositories (like Genome Expression Omnibus - GEO) we aim to identify if genes in spacial proximity have the same regulation and expression patterns. As a model study, we first analyze a K562 Hi-C experiment obtained from GEO (acession number GSE-63525) , where we: first - identified intra-cromosomal loops and which genes are in each loop, than we look for inter-cromosomal contact points between those loops. Comparing the correlations between genes locates in loops that are in contact to a random set of genes. A more consistent correlation was identified between genes in spacial proximity than random group. Moreover we identified that some genes - *e.g.* TIE1 (chr1) and TEK (chr9) – both located near the contact points between chromosomes 1 and 9 – are member of the same superpathways. This patterns indicates that spacial localization of genes may play an important role in the coordinated expression of the protein machinery necessary to a determined biological pathway. As a perspective, we also aim to identify the evolutionary patterns that leads to this intrincated spacial organization.

HD-zip classification in *Vigna unguiculata* and comparative synteny with *Phaseolus vulgaris*

Artemisa Nazaré Costa Borges, Bruna Piereck, Carolline de Jesús Pires, Flávia Tadeu de Araújo, José Ribamar Costa Ferreira-Neto, Ana Christina Brasileiro-Vidal & Ana Maria Benko-Iseppon

Universidade Federal de Pernambuco – PPGG/LGBV

Vigna unguiculata (cowpea) is an edible legume, with economic importance especially in Africa and South America. The productivity of this legume has been affected by drought, despite its higher adaptability to this stress type. Thus, it is imperative to identify stress-responsive genes associated with drought tolerance, such as the HD-zip transcription factor (TF) family that includes four subfamilies. In this study, the transcriptomes of two contrasting varieties of cowpea were analysed under water deficit (tolerant and sensitive to drought stress). TFs were identified with the iTAK program, followed by ORF-finder translation and Batch-CD-search annotation. Complete conserved domains were aligned using MEGA7 and a phenogram was built with *Neighbor-Joining* method (*bootstrap* 1000 replications). The identified HD-zip candidates were anchored at *Phaseolus vulgaris* genome through a BLASTn (*cut-off* = $1e^{-90}$), and their positions were visualized using Circos program. The iTAK program identified 88 candidates. After annotation with CD-Search, 46 were complete and employed in further analyses. As expected, the phenogram revealed four groups separating the four subfamilies of HD-zip (I - IV) indicating a conservation between subfamilies. Subfamily I (drought stress responsive) was the most abundant with 29 sequences, followed by II, III, and IV with nine, six and two sequences, respectively. The higher amount of subfamily I was expected since the transcriptome analysed was generated under water deficit. Subfamilies III and IV were more closely related, sharing the START motif close to the N-terminal region. After filtering redundancy, the anchoring in *P. vulgaris* genome reported 50 loci corresponding to the tolerant plant (*Pingo de Ouro* - PO) and 67 to the sensitive (*Santo Inácio* - SI) displaying a higher microsynteny (gene conservation) with the drought sensitive variety. All chromosomes presented HD-zip representatives, mostly close to terminal regions, being more abundant in *Pv03* and *Pv04*, and less abundant in *Pv01* and *Pv07*. Chromosomes *Pv02*, *Pv05*, and *Pv10* exhibited sequences also in the pericentromeric region, whereas *Pv09* presented sequences in the interstitial region. The minor part of the sequences was in the centromeric region (*Pv02* and *Pv03*). HD-zip sequences were displayed in clusters, with superposition and tandem duplication, indicating evolutionary divergence from a common ancestor by tandem duplications, as reported to other gene families in legumes. The identified HD-zip candidates represent valuable genetic resources and potential targets for genetic transformation of cowpea and related species.

CAPES, CNPq, FACEPE

CattleQTLdb analysis to increase understanding of the functions of milk proteins genes

Matosinho, C.G.R¹; Rosse, I.C¹; Fonseca, P.A.S¹; Assis, J.G^{1,2}; Oliveira, F.S^{1,2}; Araujo, F²; Salim, A²; Lopes, B.C³; Arbex, W.A⁴; Machado, M.A⁴; Peixoto, M.G.C.D⁴; Verneque, R.S⁴; Martins, M.F⁴; Coimbra, R.S⁵; Silva, MVGB⁴; Oliveira, G^{2,6}; Carvalho, M.R.S¹

¹ Departamento de Biologia Geral, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil, ² Grupo de Genômica e Biologia Computacional - Centro de Pesquisas René Rachou – FIOCRUZ-Minas, ³ EPAMIG – MG, ⁴ Embrapa Gado de Leite – MG, ⁵ Neurogenômica, Centro de Pesquisas René Rachou –FIOCRUZ-Minas, ⁶ Vale Technology Institute, PA

Brazil has one at the largest commercial cattle herds in the world and is the fourth largest milk producer worldwide, yielding approximately 35 billion kg of milk per year. Brazilian herds consist of taurine breeds (*Bos taurus*), indicine breeds (*Bos indicus*) and their crossbreeds. However, the genetic basis underlying the milk features and/or differences above these breeds is only partially known. In a previous study, we identified 64 SNVs and 6 INDELs in genes which codifying milk proteins from Guzerá and Gir Genome, such as α S1-casein (*CSN1S1*), α S2-casein (*CSN1S2*), β -casein (*CSN2*), κ -casein (*CSN3*), α -lactalbumin (*LALBA*), β -lactoglobulin (*LGB*) and lactotransferrin (*LTF*). However, despite we study that identified variations in these genes, it is still necessary to have a better understanding of the function of these genes in the yield of milk proteins and if these genes have function in other relevant characteristics economically also, for example analysis in QTLs. In this context, the objective of the present study was to verify *in silico*, if these genes co-locate with QTLs of economic importance deposited in Cattle QTLdb. We search for QTLs associated for *CSN1S1*, *CSN1S2*, *CSN2*, *CSN3*, *LALBA*, *LGB* and *LTF* genes in Cattle QTLdb. We select all QTLs associated with these genes found and analyzed what characteristics are related with gene. QTLs regions (60) were identified co-locating as the seven genes referred above. We found one (1.7%) in *LALBA* gene; two (3.3%) in *CSN1S2* gene; thirteen (21.7%) in *LTF* gene; fifteen (25%) in *CSN3* gene; nine (15%) in *CSN1S1* gene; and twenty (33.3%) in *CSN2* gene. In *LGB* gene, no QTLs were identified. *LTF* and *LALBA* genes co-locate with QTLs for calving interval and clinical mastitis, and fertilization rate, respectively, in addition to milk protein yield. *CSN2* and *CSN3* genes co-located with QTLs for milk alpha-lactalbumin and beta-lactoglobulin content. QTLs analysis is an important tool for understanding of the function and characteristics related with genes. These results suggest the existence of regulatory mechanisms regulating in trans the milk protein content.

Supported by: CAPES, CNPq, FAPEMIG (CBB-1181/0 and TCT 12.093/10), NIH-USA (TW007012), CAPES/CDTS-FIOCRUZ, FIOCRUZ-MG, PDTIS-FIOCRUZ - Platform RPT04B, Bioinformatics BH, Embrapa.

In silico identification of the effects of genetic variants in transcription factors recognition sites in regulatory regions of candidate genes for reproductive disorders in cattle

Diniz, LAF¹; Fonseca, PAS¹; Paiva, AE¹; Santos, FC¹; Rosse, IC¹; Moura, GS²; Santos, DJA³; Oliveira, G^{4,5}; Andrade, VJ²; Vale-Filho, VR²; Silva, MVGB³; Carvalho, MRS¹

¹ Departamento de Biologia Geral, Universidade Federal de Minas Gerais, Belo Horizonte;

² Escola de Veterinária, Universidade Federal de Minas Gerais, Belo Horizonte; ³ Embrapa Gado de Leite, Juiz de Fora; ⁴ Grupo de Genômica e Biologia Computacional, Centro de Pesquisas René Rachou (CPqRR) - FIOCRUZ, Belo Horizonte ; ⁵ Vale Technology Institute, Belém, PA, Brazil

Identify genetic variants with high functional potential among all the variants identified in a whole genome sequencing is a hard task. Variants in regulatory regions have a great functional potential in multifactorial traits. The frequency of reproductive disorders increase with the intensity of selection for production traits in bovine herds, suggesting a regulatory component shared in both processes. In the present study, we propose a pipeline for identification of variants with high functional potential in regulatory regions (5'-UTR and promoter), using as example a dataset composed by variants identified in a bull affected by reproductive disorders. To reach this aim, the variants present in the affected animal were compared with variants identified in three healthy Gir bulls. After this step, only the variations observed exclusively in the affected animal were kept for the next analyses. The exclusive variants were subjected to a functional annotation using the software NGS-SNP. The variants mapping on 5'-UTR and upstream regions were evaluated respective to their co-location with QTLs for reproductive traits. This step was developed using scripts in R and Pearl, developed in house. Three variants in 5'-UTR and six variants in upstream regions were identified in candidate genes for male reproductive traits. The selection of the variants with the highest functional impact was based on the following criteria, using these bioinformatics tools: 1) genomic context (MatInspector and NCBI); 2) evolutionary conservation (ECR Browser and Mulan); 3) alterations in transcription factor recognition sites (Mulan and MatInspector); 4) alterations in the probability of recognition of a transcription factor-binding site (Cister); 5) ratio of co-expression in male reproductive tissues of the target gene and the transcription factors (BioGPS); and, 6) involvement of the candidate gene and transcription factors in biological processes related to reproductive disorders (scientific literature). At the end of these analyses, three genetic variants with high functional potential were identified, mapped in the upstream region of three important positional and functional candidate genes for reproductive disorders in bovines (*DGAT1*, *ACTN1* and *INHBA*). Using this approach it was possible to filter variants from a whole genome sequencing (over 3 million), associate those variants with QTLs for male reproductive traits (over 180.000) and, finally, select those with the highest functional potential in important candidate genes for reproductive phenotypes (3 variants). Furthermore, suggesting that the approach proposed is efficient to identify variants with high functional potential for complex traits.

Supported by: CNPq, CAPES, FAPEMIG

Metagenomic analysis of the Southern Brazilian Atlantic Forest soil using next-generation sequencing technologies

¹Janynne Palheta, ²Michelle Zibetti Tadra-Sfeir, ²Emanuel Maltempi de Souza,
²Fábio de Oliveira Pedrosa, ^{1,3}Helisson Faoro

¹Laboratory of Bioinformatics, Professional and Technological Education Sector,
Universidade Federal do Paraná, ²Department of Biochemistry and Molecular Biology,
Universidade Federal do Paraná, ³Laboratory of Gene Expression Regulation, Carlos
Chagas Institute, Fiocruz-PR

Metagenomics allows the direct access to the DNA of the environmental bacterial communities without cultivation. Applying the Next Generation Sequencing technology (NGS) to environmental DNA has been provided precise information about the species that are present in a specific environment (microbiota) and the genes that these microorganisms are carrying (microbiome). In this work, we used the MiSeq and Ion Proton platforms to sequence the total DNA and the 16S rRNA gene from soil samples of the Southern Brazilian Atlantic Forest. The first group was formed by samples MA02, MA05 and MA07, collected in the winter of 2004 at 900, 653 and 32 meters of altitude. The second group was formed by samples MAF1, MAF2 and MAF3, collected at the same site of the first group in the summer of 2007. The total DNA from all six samples were purified using MoBio Power Soil kit and the 16S rRNA gene was amplified using universal primers customized with the Illumina adaptors sequence. The analysis of the resulting data, using QIIME package, revealed the presence of 33 bacterial phyla with the predominance of the Acidobacteria phylum (49.4%) and Proteobacteria (24.6%). The less abundant bacterial phyla were Chloroflexi (2.5%), Nitrospirae (2.2%) and Actinobacteria (1.9%). There was no alteration in the dominant or in the less represented phyla with the time and season. The total DNA was also sequenced on the MiSeq and Ion Proton platforms yielding 3.6 Gbp, 2 Gbp and 4 Gbp for samples MAF1, MAF2 and MAF3, respectively. The functional analysis on the MG-RAST server, using predicted protein sequences, based on COG groups showed that 41% of the reads were related to general metabolism followed by cellular process and signalization (22%). Based on the subsystems of MG-RAST, 11.43% of the reads were related to metabolism of carbohydrates and 8.55% to amino acids and derivatives. The reads obtained from total DNA sequencing were also submitted, separated by each platform and using a hybrid strategy, to the *de novo* assembling process through MegaHIT and CLC genomic workbench packages. The CLC assembler and MiSeq platform obtained the best results, measured by the number of contigs above 1,000 bp and the length of the larger contig: 16,426/35,630, 4,840/5,841 and 2,248/6,699 for samples MAF1, MAF2 and MAF3, respectively. All these data reflects the vast diversity of the soil, which make difficult to assemble large genomic regions without a large sequencing coverage, even using a hybrid sequencing strategy.

Inference of distant homologs in Protozoa by pHMM–pHMM comparison for the identification of superfamilies

Darueck Campos, Rodrigo Jardim, Alberto M. R. Dávila

Oswaldo Cruz Institute, Acre Federal Institute

According to World Health Organization, the major diseases in tropical countries, such as malaria, sleeping sickness, Chagas disease, leishmaniasis, amebiasis and giardiasis, are causing by protozoan parasites, which together threaten more than a quarter of the world population. In recent years, as a result of the work of several research teams, 71 Protozoa species were fully sequenced, but a majority portion of their proteins have not been functionally annotated yet. The use of pHMM (profile Hidden Markov Model) for identifying distant orthologs in those Protozoa is considered more efficient than other techniques such as comparison of protein sequences or between pHMM and protein sequences. The main reason is its potential to discover more distant homologs. Furthermore, this methodology may also contribute to the functional annotation of proteins thus enabling the improvement of knowledge about the species under study. In theory, distant homologues identification might result in the protein superfamilies identification. In light of this, we aimed to identify superfamilies by analyzing 3 Protozoan genomes: *Cryptosporidium muris*, *Entamoeba invadens* and *Trypanosoma grayi*, chosen for their evolutionary distance, using pHMM- pHMM. This methodology, was able to identify 94% of distant orthologs among all the orthologous groups inferred from the three species. Considering only 2 species, our methodology was able to identify an average of 75% of distant orthologs between *C. muris* and *T. grayi*, 50% between *C. muris* and *E. invadens* and for *T. grayi* and *E. invadens* we found 60% of distant orthologs. Our results are encouraging and allow the annotation of proteins based on distant homology inference.

Comparative Genomics between two different biovars of *Corynebacterium pseudotuberculosis* isolated in the same host

Rafael Cabús Gantois Santos¹, Thiago Jesus Sousa², Doglas Parise², Daniela Costa Arruda², Anne Cybelle Pinto Gomide², Henrique Figueiredo³, Vasco Azevedo²

¹College of Technology and Engineering, Salvador University; ²Laboratory of Cellular and Molecular Genetics, Department of General Biology, Institute of Biological Sciences, Federal University of Minas Gerais; ³National Reference Laboratory for Aquatic Animal Diseases of Ministry of Fisheries and Aquaculture, Federal University of Minas Gerais;

Corynebacterium pseudotuberculosis is classified as a Gram-positive bacteria, been responsible for a large amount of diseases around the world, like *Caseous lymphadenitis* (CLA) in goats, cattle and sheep by Ovis biovar and cattle and equines by Equi biovar. There are lots of genomes sequenced of different strains in worldwide databases like NCBI and the amount of data is rising. This work is analyzing two different strains of *Corynebacterium pseudotuberculosis*: Strain I37 and Strain CPI19, both isolated in Israel's cattle, but strain I37 has nitrate reduction (meaning it's Equi biovar) and strain I19 don't (meaning it's Ovis biovar). Both genomes were sequenced at National Reference Laboratory for Aquatic Animal Diseases of Ministry of Fisheries and Aquaculture (AQUACEN) using a 400 pb fragment library in Ion Torrent PGM™ platform and as a result deliver eight contigs with a 160.98-fold coverage in I19 and twenty-nine contigs with a 121.51-fold coverage in I37. In both genomes were performed a scaffolding process with SPAdes 3.6.0 and got assembled at Laboratory of Cellular and Molecular Genetics (LGCM) using SIMBA platform as a first step with a reference found at NCBI database, then the software CLC Workbench was used for fill the gaps. RNAmmer platform was used to identify the RNA genes and RAST platform was used as a automatic annotation. UNIPROT database was used to search sequences and the software Artemis was used to manually curate as a last step of the modeling process in both genomes. For the comparison, the genomes run in BRIG (two times, one with each as a reference and the other as a comparison object) and MAUVE, with both results it was possible to find where are the differences between them and investigate these spaces searching for proteins (as a CDS product) that are only in one genome. Some of these products was discarded of this work because they are hypothetical proteins and need to be studied harder to find out what they do, this work meant to compare known proteins. Now only eight known strain I19's exclusive proteins and twenty-six known strain I37's exclusive proteins left, but five of the strain I37's exclusive proteins belongs to Nar operon (alpha, beta, gamma, delta and the nitrate transporter), the responsible to determinate the Equi biovar. All the other proteins are being investigated for pathogenesis-related in the literature.

Relative Evaluation of NoSQL Databases For Manipulating Genotype Data

Vinícius Junqueira Schettino¹, Arthur Lorenzi Almeida¹, Fernanda Nascimento Almeida^{1,2}, Wagner Arbex^{1,2}

¹*Federal University of Juiz de Fora (UFJF)*, ²*Brazilian Agricultural Research Corporation (Embrapa)*

One of the greatest challenges on bioinformatics research is to manipulate the data. Genotype files, vastly used in this field, are known by their high dimensionality and unbalancing. These aspects are some of the reasons RDBMSs, traditionally signed for tabular information persistence, have not been shown as good infrastructure to analysis that rely on this kind of data. Therefore, this abstract aims to evaluate the relative performance among NoSQL engines on genotype data manipulation. In this text we present the extension of previous studies, encompassing the viewpoint of scalability, as well as including results from three representatives of distinct NoSQL databases families. For our evaluation, we used the Yahoo! Cloud Server Benchmark, a framework designed for asserting NoSQL databases aspects. Three databases were considered, each of them representing one family of NoSQL engines: Tarantool as "Key/Value Based", MongoDB as "Document Based" and OrientDB as "Graph Based". We simulated two populations with 5,000 individuals, with a hypothetical SNP sequence for each individual. One population with 20,000 SNP markers, the other with 56,000. Two scenarios were considered: One with 5,000 insert operations, and another with 10,000 equally divided read and update operations. To assert the scalability, we measured the throughput of each workload for these engines. On the insert scenario, using 20,000 SNP markers, MongoDB was capable of handling 56.8 ops/s on average, followed by Tarantool and OrientDB with 41.8 and 33.7 ops/s, respectively. For the 56,000 SNP markers population, MongoDB, Tarantool and OrientDB handled, on average, 21.7, 14.0 and 12.0 ops/s each. For read and update operations with the 20,000 SNP markers population, Tarantool executed on average 515.1 ops/s, followed by MongoDB with 170 ops/s and OrientDB with 26.0 ops/s. With 56,000 SNP markers, the results were: 307.5, 50.0 and 9.9 ops/s on average for Tarantool, MongoDB and OrientDB, respectively. Comparing these results, Tarantool was capable of keeping an average 33.6% of its performance on insert operations and 59.7% on reading/updating when we increased the SNP markers sequences from 20,000 to 56,000. MongoDB kept on average 38.2% on insert operations and 23.4% on read/update operations, while OrientDB preserved 35.5% of its performance on insert operations and 38.3% on read/update operations. For insert operations, the three engines scaled similarly, though MongoDB did the best absolute throughput for this kind of operation. Tarantool scaled better for read and update operations and presented the best absolute throughput.

Supported by CAPES, CNPq, Embrapa, FAPEMIG and UFJF.

Metagenomics insights reveals functional patterns among soil microbial communities of global biomes

Melline Fontes Noronha¹, Gileno Vieira Lacerda Junior¹, Jack A. Gilbert ^{2,3} and Valéria Maia de Oliveira¹

1 Microbial Resources Division, Research Center for Chemistry, Biology and Agriculture (CPQBA), University of Campinas, 2 The Microbiome Center, Department of Surgery, University of Chicago, Chicago, IL, USA and 3 The Microbiome Center, Bioscience Division, Argonne National Laboratory, Lemont, IL, USA

A biome is a geographical unit characterized according to its vegetation type, macroclimate, soil, and specific elevation. In contrast, a microbiome is a mix of microorganisms that coexist in a defined space. Although soil microbial communities have shown to vary across many spatial scales, soils between ecosystems showed to be leading by some biogeographical trends. In order to investigate functional convergence within soils from the same biome type, thirty publically-available metagenomes from 11 globally distributed biomes were selected and clustered by biome groups (i.e. forest, grasslands, tundra, semiarid and desert) based on vegetation features. Functional analyses revealed a close pattern among biomes groups, in which DNA repair, central carbohydrate metabolism, and antibiotic resistance were the most statistically different metabolism annotated by SEED subsystems among biome groups. In order to provide a better analytical resolution of those metabolisms, metagenomic reads were annotated using the Carbohydrate-Active enZYmes database (Cazy), Antibiotic Resistance gene DataBase (ARDB) and, additionally, the Heat Shock Protein Information Resource (HSPIR). Carbohydrate-active enzyme analyses showed that biomass degradation, sucrose and starch metabolism, cell wall biosynthesis and alginate degradation were overrepresented in forest and grasslands soils. As expected, desiccation and other stress resistance genes were more abundant in deserts and semiarid soils. Antibiotic Resistance Genes (ARGs) were prevalent in forest and grassland soils, where multidrug efflux pumps were the most abundant ARG class, with the majority of the reads assigned to Proteobacteria. Heat Shock Proteins (HSPs) were more abundant in tundra, semiarid and desert soils. Although HSP70 and HSP100 were uniformly distributed across biomes, while HSP60 and HSP20, which are predominantly from the Archaea, were more abundant in the Saline Desert soils. Our results suggest that local environmental conditions select for the enrichment of specific functions important for survival in those ecosystems.

Complete genome sequence of *Corynebacterium pseudotuberculosis* 33

Viana, MVC¹; Parise, D¹; Sousa, TJ¹; Benevides, LJ¹; Mariano, D¹; Rocha, FS¹; Bagano, P¹; Guimaraes, LC²; Pereira, FL¹; Dorella, FA¹; Rammes, R²; Silva, A²; Selim, SAK³; Salaheldean, M³; Figueiredo, H¹ and Azevedo, V¹

¹Universidade Federal de Minas Gerais, UFMG, Belo Horizonte, Minas Gerais, MG;

²Universidade Federal do Pará, UFPA, Belém, PA; ³Faculty of Veterinary Medicine, Cairo University, Cairo, Giza, Egypt

Corynebacterium pseudotuberculosis biovar Equi is a Gram-positive, pleomorphic, facultative intracellular pathogen that causes outbreaks of Oedematous Skin Disease (OSD) in bu alo from Egypt. Isolates from this host harbor the diphtheria toxin gene. Sequencing of *C. pseudotuberculosis* genomes isolated from this host will help to understand the pathogenicity of the species, including whether the diphtheria toxin is required for infection. *C. pseudotuberculosis* 33 was isolated from a bu alo diagnosed with OSD in Egypt, in 2008. The genome was sequenced by Ion Torrent platform, using Ion PGM Template OT2 400 Kit and Ion PGM Hi-Q Sequencing Kit. The sequencing resulted in a number of 1,439,326 reads and a total of 250,725,445 bases that were checked for quality by FastQC v0.10.1. A *de novo* assembly was done with Newbler v2.9 and generated 14 contigs with a N50 of 536,189. Scaffolding was done by CONTIGuator v2.7, using *C. pseudotuberculosis* 31 (CP003421.2) as the reference genome. An in-house script established the *dnaA* gene as the beginning of the chromosome. These steps have been integrated in the webtool Simba v1.2.1. The gaps were closed by a reference assembly using CLC Genomics Workbench v6.5 and *C. pseudotuberculosis* 31 as the reference. The genome was annotated using RASTtk. Sequencing errors causing frameshifts were curated by checking for indel errors in the reads, using CLC Genomics Workbench v6.5 and Artemis v16.0.0. The assembled genome has 2,403,550 bp, 52.08 % of GC content, and a mean sequencing coverage depth of 104.11x. The annotation has 2,281 CDSs, 52 tRNAs, and 12 rRNAs. An insertion of 40Kb containing the diphtheria toxin was found, as in *C. pseudotuberculosis* 31. This genome and other bu alo isolates will be used in a comparative genomic study to better understand the pathogenic mechanisms of OSD.

Financial Support: CAPES, CNPq and FAPEMIG

Evolution of heterotrophy: Genes needed for regulation of the acidity of pancreatic juice appeared recently in man evolution.

Fenícia Brito Santos, Carlos Alberto Xavier Gonçalves, José Miguel Ortega

Laboratório de Biodados, Instituto de Ciências Biológicas, UFMG.

The pancreas works as two glands with endocrine and exocrine functions. The exocrine function of the pancreas consists of the acinar cells that are responsible for synthesize and secrete digestive enzymes comprising the pancreatic juice. The digestive enzymes - namely amylases, proteases and lipases - are stored Zymogen granules. The stimulation of acinar cells by acetylcholine and cholecystokinin by the Parasympathetic nerve and the intestinal I-cells, respectively, triggers the generation of an intracellular Ca^{2+} signal cascade. This signalization culminates on the secretion of the enzymes in the zymogen granule. In general, the proteins related with the pancreas enzymatic secretion are associated with the zymogen granule migration and fusion on the acinar apical membrane, releasing the enzymes on the lumen, or adjusting the pH and ion equilibrium in the cell. We determined the Lowest Common Ancestor (LCA) for the genes on this system to investigate their origin along the evolution. The pancreas main secretive function is found in the acinar cells. In these cells there are proteins responsible for receiving the external stimulation from the secretagogues – M3, CCKAR – and for the pH regulation and the Ca^{2+} equilibrium - PLC, CD38/157, SOC - that have their origin in Bilateria. These proteins are involved on key processes such as the zymogen granule migration. Other proteins, remarkably some transmembrane transporters, are a recent acquisition of these cells emerging in more recent clade such as Euteleostomi, although some of them have ancient functions, such as SOC, ATP, NHE1 and AE2. Similarly, although the function for the secreted enzymes is rather ancient some of their sequences show a remarkably recent origin such as CEL in Euteleostomi and PLA2 in Amniota. Thus it is reasonable to suggest that the central components of the pancreatic secretion system emerged in Bilateria. The other type of cells found in the pancreas are the duct cells, whose main function is the secretion of bicarbonate, which neutralize the acidity of gastric contents. Several of the proteins found in these cells have more recent origins, in clades such as Gnathostomata, Euteleostomi and Dipnotetrapodomorpha. Although the function of the duct cells are not critic for the digestive process, they are essential for the occurrence of this process, otherwise the pancreatic juice would digest the intestine itself. Thus, we conclude that this important gland, deeply associated with the heterotrophy needs recent biological functions for its systemic functioning, originated between the clades Bilateria and Euarchontoglires.

Walking through old routes to reach new destinations: unraveling the origin of the mammary gland

Lissur Azevedo Orsine¹, Elisa Rennó Donnard Moreira², José Miguel Ortega¹

¹*Biodata Lab, Federal University of Minas Gerais, Brazil*, ²*Garber Lab, University of Massachusetts Medical School, United States of America*

The mammary gland is closely related to the evolutionary success of Mammals. It is responsible for pup's nutrition and immunity at the beginning of pup's life, which gives it advantage in the fight for survival. However, despite the significance of the mammary gland to Mammals, the understanding about the molecular mechanisms which control the development of this organ is incipient. And, regarding to its origin and evolution, the lack is still bigger. Thus, the aim of this work was to estimate the origin of the genes involved in the development of the mammary gland and, then, estimate the origin of the mammary gland. Therefore, the first step consisted in collecting information on the relevant genes (and the interactions between them), through text-mining tools, and assembling a pathway diagram and a description for each phase of the mammary gland development (embryonic development, puberty, pregnancy & lactation and involution). Subsequently, the origin of each gene was estimated by determining the lowest common ancestor of the organisms containing copies of each gene, discovered with homology-clustering tools. The origin of the system was inferred based on the origin of the genes. Four pathways were generated in this work, one for each stage of mammary gland development, and they are accompanied by their descriptions. In total, 310 genes and 795 biointeractions were found. With respect to evolutionary origin, genes were found to be present since cellular organisms up to Boreoeutheria, and 80-97% of genes were already present in fishes (among Gnathostomata, Teleostomi and Euteleostomi). The comparison between subpathways revealed that embryonic development and puberty involve genes with predicted origin up to Tetrapoda-Amniota, while pregnancy & lactation and involution comprise genes with predicted origin up to Mammalia-Eutheria-Boreoeutheria. A common process along the history of life is co-option, i. e., the recruitment of pre-existing pathways to generate new structures and functions. This seems to be the case of the mammary gland, since the genetic potential to generate this organ had already existed long before the Mammals' origin. That is evolution walking through old routes to reach new destinations.

Mosquitoes Mobilome

Elverson Soares de Melo, Gabriel da Luz Wallau

Aggeu Magalhães Research Center (CPqAM) - Fiocruz Pernambuco

Genomes of living organisms are composed of stable (genes) and unstable components (mobile genetic elements or transposable elements(TEs)). In the last years an increasing amount of evidence highlights the importance of TEs as major players in the genome evolution providing raw material for natural selection. Such elements have been shown to reshape the host genomes in a multitude of ways as generating chromosome rearrangements, rewriting of transcription networks and also being co-opted for a new advantage features for the host species. Although much is known about TEs, such knowledge is restrict to few model organisms, for most of currently available genomes poor information exists even at the mobilome characterization level which reflects the almost absence of data about TE impact on these genomes. Mosquito genomes are available for some time from now and its mobilome annotation has both extremes: *Anopheles gambiae* genome with well characterized mobilome, *Aedes aegypti* and *Culex quinquefasciatus* with some degree of characterization and the other new 16 Anophelinae genomes available in 2016 with only partial description of TEs. Based on the reasoning presented above the main goal of this study was to evaluate the TE impact on mosquito genomes by re annotating it using recently developed and powerful pipeline for mobilome characterization making use of information from mosquitoes genomes and transcriptomes. REPET package was used for genome-wide TE characterization in seven genomes so far. Our preliminary results shows that this pipeline detected a higher TE content (20%) in the *Anopheles gambiae* genome than previously reported in the literature (17%). A different TE proportion also was found for all other 6 species: *A. arabiensis* (9,59%), *A. coluzzii* (10,41%), *A. merus* (9,6%), *A. melas* (4,71%), *A. epiroticus* (5,51%) and *A. christyi* (0,81%). In addition, we also performed a deeper analysis for each TE superfamily and could detect a different dynamics of those superfamilies both inside of each genome as well as among them. Some families being old components of the genome probably not transposing anymore and other probably highly active families which may have a deeper impact on a shorter evolutionary scale. In summary, our data shows the need of standardized and deeper mobilome characterization of the mosquitoes genomes and that TEs experienced different expansion time and extent inside of each genome. Moreover mosquito mobilome are evolving under different selection constraint among those related species studied.

B-cell epitopes prediction in trypanosomatids genome core

Anderson Coqueiro Santos¹, Leandro Martins de Freitas²

¹Institute of Biological Sciences at UFMG and ²Multidisciplinary Institute of Health at UFBA

The trypanosomatidae family is one part of the phylum euglenozoa including species such as *Leishmania* and *Trypanosoma spp.* There are some potent vector-borne diseases for humans, and others mammals, mainly in subtropical and tropical countries. Leishmaniasis, Chagas diseases, and human African trypanosomiasis (HAT) are some pathologies infecting people worldwide, responsible for approximately 9 million people infected in all world. The observation of orthologous - shared genes present in a common ancestor among species - or paralogous genes - duplicated shared genes - has been one of the main research conducted. All genes shared by different groups of organisms are defined as "genome core". In our study, Trypanosomatids orthologous proteins groups were retrieved from OrthoMCL database. These proteins are conserved in eukaryotic organisms and normally two paralogs sequences were found, as observed in this work for *T. cruzi*, *T. congolense* and *T. brucei*. The orthologous group were submitted to predictions using Bepipred software to find possible B-cell epitopes using a threshold for epitope assignment greater than or equal to 0.35. The statistical analyses were performed using R software (p-value <= 0.05). It was considered as predicted b-cell epitopes each amino acid assigned by the software. The predictions for b-cell resulted in a total 5,865,620 possible epitopes (each single amino acid) with an average of 243 epitopes by proteins. A total of 177,946 b-cell predicted regions (more than 5 amino acids grouped linearly). The regions showed variation in amino acid length, ranging from 5 to 428. Analyzing the proportion of predicted amino acids over the total protein length, we observed that *Leishmania* proteins have a larger number of epitopes when compared with *Trypanosoma* proteins. Some *Leishmania* proteins have more than 80% constituted by b-cell predicted epitopes. When we compared the number and length of b-cell epitopes regions, the *Trypanosoma* species has lower number compared with *Leishmania* species with statistical significance. We also investigated epitopes position and conservation in the aligned sequences in each orthologous group and it was found they are conserved. With this work we observed that *Leishmania* species presented linear epitopes length greater than *Trypanosoma* species. These results point the importance of immune response against these parasites, while *Leishmania* has an obligatory intracellular phase, not being accessible to b-cell, *Trypanosoma* exposed more time during infection could evolve to less b-cell regions. Which may lead to selection of trypanosomes that has a lower number and length of epitopes for b-cell explaining the epitope differences among species.

Construction of alternative algorithm for development of phylogenetic trees using InterPro entries and frequency vectors of tri-peptide

Lucas Felipe Silva¹, Dalila Dominique Duarte Rocha², Lara Maria Silva Miranda¹,
Matheus Allef Cruz¹, Thiago do Carmo Librelon Rocha¹, Bráulio Roberto
Gonçalves Marinho Couto¹, Marcos Augusto dos Santos³

¹*Centro Universitário de Belo Horizonte (Unibh);* ²*Universidade Federal de São João del-Rei;* ³*Universidade Federal de Minas Gerais (UFMG)*

In the study of the evolution of species, the use of phylogenetic trees to verify the relationship among them is essential. However, the evolutionary reconstruction organisms using traditional phylogenetic methods can be affected by errors, e.g., misalignments or by using a limited number of genes. In addition, alignment methods of complete sequences of genomes are impossible because they demand a huge computational effort. In this context, representation of proteins as vectors in multidimensional space opens up possibilities for the application of linear algebra methods to investigate such relationships. The InterPro database integrates predictive models or "signatures" of proteins, describing it as a tool in the study of evolutionary processes. This research looks for answers to the following questions: a) genomes analyzed by linear algebra methods, using proteins as vectors of the frequency of tripeptide and InterPro entries can generate valid phylogenetic relationships from a biological viewpoint? b) What is the computational performance of linear algebra techniques when used to generate phylogenetic trees, compared to classical methods? Two sets of data were used: complete genomes were analyzed from 14 species of plant models, and 317 complete Eukaryotic genomes, retrieved from the UniProt database (<http://www.uniprot.org/proteomes/>). Classical methodology (pair-to-pair alignments) and linear algebra technique with tri-peptide vectors and frequency vectors of InterPro entries were applied to both datasets. In classical analysis, the sequences of the genomes were tested in the programs: MEGA, ClustalW, Clustal Omega, MUSCLE, BioEdit, and CLC Sequence. Needleman-Wunsch global alignment algorithm was used to generate data necessary to build classical trees. Unfortunately, there is no valid results due to computer problems and none of the programs submitted supported the large amount of data (for example, in the plant models data, evaluated genomes had, on average, approximately 13 million amino acids). To use linear algebra methods, we developed a matrix with the presence (1) or absence (0) of InterPro entries for each complete genome analyzed. In another matrix representation, all protein sequences of each complete genome were transformed into vectors of tri-peptides frequencies. By using Linear Algebra algorithm, it was possible to construct phylogenetic trees that showed similar results for both representation vectors, tri-peptides and InterPro entries. It has been found that the distribution of species in the dendograms was generated according to the taxonomy presented in the literature. The results showed that genomes can be evaluated using linear algebra techniques.

Biological modules associated with prophage density in pathogenic and commensal *Escherichia coli*

Tarcisio José Domingos Coutinho¹, Glória Regina Franco² e Francisco Pereira Lobo¹

1 - Departamento de Biologia Geral – ICB/UFMG, 2 - Departamento de Bioquímica e Imunologia – ICB/UFMG

Integrated phages (prophages) play an important role in the genomic diversification and fitness cost to the infected host, since they are major contributors to the diversity of bacterial gene repertoires but may also lead to host death. Prophages able to propagate horizontally may allow bacteria to adapt to many ecological niches through horizontal transfer of biological modules. Our evolutionary interest in prophages consist in understanding the role of adaptive genes that reach bacterial genomes through phage integration and their contributions to the complex antagonistic and mutualistic prophage-bacteria interactions. We used comparative genomics and statistical analyses to study the influence of genes carried by prophages in *Escherichia coli* pathogenic and commensal lineages. We downloaded 33 and 17 complete genomes of pathogenic and non-pathogenic *E. coli* strains, respectively, from the National Center for Biotechnology Information (NCBI). We used PHASTER (<http://phaster.ca/>) to analyze bacterial genomes in order to find prophages (intact, incomplete or questionable), Rstudio (<https://www.rstudio.com/>) for graphics and KOMODO2 (<https://www.komodo.cnptia.embrapa.br/>) for correlation analyses. We detected 470 phages in *E. coli* genomes, 359 in pathogenic and 111 in non-pathogenic lineages. For each genome, we calculated the phage density (number of phages divided by genome length). We observed that pathogenic *E. coli* contain a significantly higher number of prophages when compared with non-pathogenic strains (Wilcoxon test, p-value = 0.009283). In order to detect potential modules associated with phage density, we searched for Gene Ontology (GO) terms whose frequency increases or decreases with prophage density in the two groups of *E. coli* analyzed. We selected 225 GO terms (115/110 terms with Pearson correlation > 0.5 and < -0.5 respectively) in pathogenic lineages and compared their correlation values with the ones found in non-pathogenic lineages. We found several of these correlated terms to be exclusive of pathogenic lineages (e.g. catalase, urea, hemolysis, parasitism, urease). Other terms presented a correlation higher than 0.5 in pathogenic lineages, but virtually no correlation in non-pathogenic ones that are related to metabolic pathways and symbiosis (e.g. nickel cation binding, cysteine-type peptidase activity, metallochaperone activity, DNA replication initiation, modification by symbiont of host morphology or physiology). Together, our analyses suggest that phages carry several biological modules that favor a pathogenic phenotype in *E. coli* lineages, and also others that may affect their metabolism, with both classes favoring bacterial fitness and evolution.

Inferring the demographic history of *Cnesterodon brevirostratus* using bioinformatics tools

Quinsani A.D., Freitas M.V., Ramos-Fregonezi A.M.C., Malabarba L.R., Fagundes N.J.R

Genetic department, Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, Brazil

General Biology Department, Universidade Federal de Viçosa, Viçosa, MG, Brazil

Zoology Departament, UFRGS, Porto Alegre, RS, Brazil

Cnesterodon brevirostratus is a freshwater fish species occurring in highland grasslands in South Brazil. Recent genomic data have shown that this single taxon consists of four evolutionary different lineages. Some of these lineages may be found in the same river basins, sometimes syntopically, and the different lineages show morphological differences in the gonopodium, which can be related to sexual selection. However, details of its divergence and evolutionary history are still elusive. In this work we used genomic (RADSeq) data to fit alternative demographic models using the allele frequency spectrum (AFS) in the dadi software. We used 6841 SNPs from 133 individuals to construct the folded AFS for each lineage. We tested five models for each lineage: neutral (constant population size), growth (exponential size change), two epoch (an instantaneous size change), bottleneck (instantaneous size change followed by exponential growth) and three epoch (two size changes, the first during a specific period and the second beginning some time ago in the past). The best-fitting model was chosen based on the Akaike Information Criterion. Preliminary results show that the best model for lineages 1, 2 and 3 was the 'growth' model, while for lineage 4 was the 'two epoch'. For lineages 1 and 2, growth ratio was similar, from about 3,000 to 30,000 individuals (95% Confidence Interval ~ 20,000 – 60,000 individuals) starting around 10,000 years ago (95%CI ~2,500 – 20,000 years ago). Despite the broad confidence intervals, results for lineage 3 suggest an older and stronger growth, arriving at a population of ~41,000 individuals (95%CI ~3,000k – 130,000 individuals) since ~20,000 years ago (95%CI 500 – 68,000 years ago), coinciding with the Last Glacial Maximum. Finally, the best-fit model for lineage 4 suggests that by 13,000 years ago (95%CI 1,500 – 41,000 years ago) a small population of 5,000 individuals (95%CI 3,000 – 7,000 individuals) grows to 17,000 individuals (95%CI 5,000 – 38,000 individuals). These results suggest that Pleistocene glacial cycles were probably the drivers of population size change for these lineages, especially related to the colder and drier glacial periods, which favored the expansion of the grassland environments in which these lineages can be found. The next step is to analyze the joint allele frequency spectrum of two lineages to gain insights into their splitting times, and to test if divergence was followed by gene flow or not. Financial support: CAPES, CNPq, UFRGS.

Improving the supertree approach by analyzing protein clusters with paralogs and including distance data

Sakamoto, T.¹, Ortega, J.M.¹

¹*Laboratório de Biodados, Departamento de Bioquímica e de Imunologia, ICB, Universidade Federal de Minas Gerais (UFMG).*

In phylogenomics, we can estimate a species tree using the supertree approach that consists in reconciling several gene trees to construct one concise tree. Currently, there are several methods that combine the information of thousand trees to construct a species tree, but some common restrictions and limitations are imposed by them. One of them is the non-acceptance of input trees containing paralogs in the analysis. This restriction is becoming more significant as the number of organisms with genomic data increases since this consequently increases the number of protein clusters with paralogy relationship and reduces the number of gene clusters acceptable for this analysis. Another key challenge for supertree approach is the estimation of evolutionary distance to allow users to address questions about the evolutionary rates and dates in the final tree. Here we present HyperTriplets, a phylogenomics tool that uses a triplet-based approach for supertree reconstruction. HyperTriplets takes as input a set of rooted phylogenetic trees in Newick format and decomposes them in triplets. The algorithm counts the occurrence of each type of triplets and uses these data to reconstruct the final tree. HyperTriplets deals with trees containing paralogs by classifying all internal nodes between speciation and duplication nodes. This allows the algorithm to extract during the tree decomposition only those triplets without duplication nodes. In the same time, HyperTriplets extracts the phylogenetic distance between pair of samples in the tree. These values are used by the algorithm to estimate the branch lengths of the final tree. HyperTriplets was tested using phylogenetic trees deposited in PhylomeDB database. The source code of HyperTriplets can be accessed at biodados.icb.ufmg.br/hypertriplets.

Financial Support: FAPEMIG, CAPES.

Reconstructing ancestral protein-protein interactions of virus-host systems

Anderson F. de Brito, John W. Pinney

Theoretical Systems Biology Group, Division of Molecular Biosciences, Imperial College London

Over the last decades of advances in biomolecular research, large amounts of biological data have been made available, which now allows us to apply integrative approaches to combine information from different levels of complexity. Such approach has proven to be of particular interest to broaden our understanding on how host-pathogen systems evolve, in particular by integrating genomic, proteomic, gene ontology, structural, and taxonomy data. In our research we have been using computational tools to infer the phylogenetic history of Protein-Protein Interactions (PPIs) between viruses and their hosts. As a starting point, a reference structure of a herpesvirus-human protein complex was taken from PDB. Searches for homologous proteins in taxonomically related species were performed in order to create multiple sequence alignments (MSAs) depicting the amino acid diversity of both viruses and hosts. Such alignments were then used to create genealogies of the protein families, yielding trees whose internal nodes represent one or more ancestral states of the existing proteins included in the initial alignments. By applying in-house methods and Maximum Likelihood approaches implemented in PAML and FastML, distributions of ancestral sequences from viruses and hosts were inferred by using the genealogies and MSAs previously mentioned. In the near future homology modeling will be used to reconstruct ancestral virus-host complexes, and variations in free energy ($\Delta\Delta G$) between existing PPIs and their ancestral states will be calculated, revealing important aspects of PPI evolution. With these results we aim to expand the understanding on how mutations (substitutions and indels) determine protein affinity in similar protein pairs, which although homologous, show remarkable differences in terms of binding energy. In addition, we intend to apply this approach to predict new PPIs, what will allow us to take advantage of the knowledge obtained in widely studied systems to better understand the protein interactions in neglected virus-host pairs.

Acknowledgments: we thank Capes and Imperial College London for the financial support.

Study of the origin of the genes controlling flower development

Castro, Beatriz M.K.; Gonçalves, Carlos A.X.; Ortega, J. Miguel

Laboratório de Biodados, Universidade Federal de Minas Gerais

Flowers are recent innovations in the evolutionary history of plants. The only extant plants that develop flowers are the angiosperms. They constitute the most diverse and cosmopolite group of plants, due to their great evolutionary success. Their main evolutive strategies are (i) the pollen dispersion by the male parts, (ii) the protection of the ovules by the female parts and (iii) the envelopment of the seeds by the fruit. In order for the flower to be formed, a complex regulatory network with different genes controls the floral development. Most of these genes belong to the MADS-box family. The main controller of this network is the LEAFY gene. During evolution, new genes can arise and enable new molecular activities, thus allowing for the development of more complex regulatory systems. This work had the objectives of (i) constructing a pathway to describe the floral development of *Arabidopsis thaliana*, a model species of the angiosperm group; (ii) grouping their orthologues with SeedServer software and (iii) investigate the evolutionary origin of the genes present on this system, by determining the Lowest Common Ancestor (LCA), clade which originated all current species carrying each gene, using Genesis webtool. The results demonstrated that the first MADS-box to appear during evolution was the AP3 gene, on origin of the eukaryotes. Other early genes of this family were AP2 and CO, which originated on the green plants (Viridiplantae). LEAFY only appeared later, among the green algae (Streptophyta). SEP1, SEP2, SEP4, AG and API all arose on the land plants (Embryophyta), whereas SEP3 and SPV only originated recently, on the clade of the vascular plants with seeds (Spermatophyta). These results show that the genes that comprise the complex regulatory network for the floral development appeared on different clades during the evolutionary history of the plants, with the main regulator of the system originating among aquatic organisms. However, most of the MADS-box genes appeared during the conquest of the terrestrial environment by the plants, along with their independence of water for fecundation and the development of the seed for the protection of the embryo. Other genes involved in the flower formation pathway, are only found on the clade of the angiosperms (Magnoliophyta) and non-basal angiosperms (Mesangiospermae). Thus, flowering comprises genes originated along all clades of plant evolution, but depending on some genes as recent as the origin of the flower.

Secondary structure changes according to evolutionary age

Ricardo Assunção Vialle, José Miguel Ortega

Departamento de Bioquímica e Imunologia, Instituto de Ciências Biológicas (ICB), Universidade Federal de Minas Gerais (UFMG), Brasil.

The origin of new genes is an important factor for innovation in all organisms. For a long time was thought that all modern genes forms were derived from other genes through the processes of duplication-divergence and horizontal gene transfer. Today, several other mechanisms are known to be involved in gene origin, such as exon shuffling, retroposition, mobile elements, lateral transfer and *de novo* origination. The *de novo* origin term refers to when a non-coding region start to code a new gene. This concept of “orphan genes” is based on comparative genomics findings showing that about 10-30% of all genes in a species show no similarity to existing proteins. In this study we analyzed 66 reference proteomes managed by the Quest for Orthologs (QfO) consortium and, estimating their age through orthology inference algorithms, we analyzed the differences of structural distributions between these groups. We tried to answer the question: Sequence independent properties are different depending on the evolutionary age of the proteins? Using information of secondary structure prediction, we noticed the presence of less structured sequences in eukaryotes than in prokaryotes. Furthermore, we found a decrease of structure in newer sequences in eukaryotes, while the orthologous proteins in prokaryotes have similar distribution parameters. We also explored the secondary structure of noncoding sequences and found that the source of new genes is compatible with the readily achieved by translation of these sequences. Perhaps the origin of secondary structures is the simple reflection of the individual potential contribution of each type of residue. Thus, the secondary structure content achieved by translation or intergenic antisense sequences is compatible with that present in said orphan genes or *de novo*.

ProteinWorldDB in a multidimensional representation of the Network of Life

Edson Machado¹, Marcos Catanho¹, Paulo Carvalho² and Wim Degrave¹

¹*Oswaldo Cruz Institute, FIOCRUZ, Rio de Janeiro, Brazil*

²*Carlos Chagas Institute, FIOCRUZ, Curitiba, Brazil*

Inter-genomic distance estimations reflect genome complexities and divergence, but are not yet well understood. Calculations of such distances with the purpose of constructing evolutionary relationships are based on the selection of a subset of homologous sequences (orthologs) and provide understanding of evolutionary relationships between genes, proteins and species, with the phylogenetic tree being a primary tool in analysis and visualization. However, inferring the "true tree" is fundamentally a difficult problem, and the traditional two-dimensional visualization is hard to interpret when a larger amount of organisms is involved. In addition, this approach ignores most of the unique traits and dissimilar aspects of the genome organization and coding/non-coding potential of organisms, paralogs, putative proteins of unknown function, etc. One can thus express inter-genomic distances comparing overall nucleotide sequences, which is not uniform due to a high variability in genetic/genomic complexities and GC bias, or by comparing the complete predicted protein set for each genome. We adapted a genomic distance method, originally based on protein similarities scores measured in bidirectional comparisons with BLAST, to use protein similarities scores measured in unidirectional comparisons using Smith-Waterman algorithm, with SSEARCH program, to infer distances between genomes and construct a distance matrix. As an example, we used data of ProteinWorldDB to apply our genomic distance method to infer distances between 210 species (117 bacterias, 49 eukaryotes and 44 archaeas) in order to construct an initial representation of the Network of Life. ProteinWorldDB stores the results of the "Uncovering Genome Mysteries" project, which examined close to 200 million predicted protein sequences from a wide variety of life forms. Those protein sequences were compared against each other through the IBM World Community Grid with the SSEARCH program to assess their similarity. This represents about 20 quadrillion (2×10^{16}) comparisons and the total computation time is projected to take the equivalent of one computer running continuously for 40,000 years. In addition, the results of the thus measured distances between the 210 species under analysis were used to construct a three-dimensional representation of the Network of Life. Support: Fiocruz, CNPq, IBM.

GO-Genesis: finding the origin of biological processes and molecular functions from Gene Ontology

Carlos Gonçalves¹, J.M. Ortega¹

¹*Laboratório de Biodados, Departamento de Bioquímica e Imunologia, ICB, UFMG*

Gene Ontology (GO) is a database comprised of terms that can be annotated to proteins to describe biological information. These terms are hierarchically organized in three distinct ontologies (Biological Process, Molecular Function and Cellular Component), with more generic terms being parents of more specific ones. Biological Processes are complex systems that require the participation of several gene products, such as “photosynthesis” or “innate immune system”; Molecular Functions are sequence-related properties of the genes, such as being able to catalyze a reaction (i.e., “phosphatase activity”) or being able to interact with another substance (like “calcium ion binding”); Cellular Components describe regions within the cell or its periphery where the gene product can be located, such as “cell membrane”, “nucleus” or “synapse”. Several of these biological processes and molecular functions are clearly very ancient, existing since the origin of life, whereas others have to be more recent innovations on the course of evolution, so we decided to estimate the moment of origin for all of them. For each GO term, we determined all organisms that had at least one protein annotated to that particular term; with that, we calculated the lowest common ancestor (LCA) for those organisms – that is, the clade on which the function or the process itself originated, at least according to the existing Gene Ontology annotations. Given the fact that each individual term had its origin independently determined, there were some cases on which a given term could be dated more recently than one of its children; since, by logic, this is an invalid result, we corrected the LCA of those ancestor terms to match the origin of the oldest of their children terms. Overall, most of the terms (over 53%) of molecular functions existing on the *Homo sapiens* are very ancient, dating back to the origin of the cellular organisms, with an expressive number also appearing on the ancient clades of Eukaryota (13%) and Opisthokonta (4%). More recently, peaks of origin of terms are observed on the clades of Bilateria (7%), Euteleostomi (4%) and Amniota (10%). As for the biological processes, about 23% of them originated on the cellular organisms, with several others also appearing on the early clades of Eukaryota (13%) and Opisthokonta (6%). Surprisingly, the clade which had most processes being originated was Amniota (26%), a rather recent one. The origin of the GO terms is available for consultation at <http://biodados.icb.ufmg.br/GO-Genesis>.

Financial support: CAPES

mtDNA Data Mining: A Global Analysis

Camilla Reginatto De Pierri¹, Bruno Thiago de Lima Nichio¹, Tetsu Sakamoto²,
José Miguel Ortega², Mauro Antônio Alves Castro¹, Roberto Tadeu Raitz¹

Federal University of Paraná¹, Federal University of Minas Gerais²

Phylogenetic methods based on comparisons of whole genomes features are relevant to provide information of many species and genes, besides the understanding of the evolutionary relationships. The origin of mitochondria and the time of acquisition in the course of evolution are some of the targets in the current research have led to many points of view by evolutionary biologists. The lack of eukaryotic intermediaries complicates the analysis of the actual mitochondrial evolutionary history. Despite several studies conducted until now and the large amount of information already known about the last common mitochondrial ancestor, we still need more evidences about the time of the organelle acquisition. In order to clarify facts not explained yet, this research was conducted in two steps. We've used a data mining strategy to explore all aspects of mitochondrial DNA and then inferred a phylogenetic tree based on pairwise distance. For this, we used RefSeq data from 6,811 organisms, totalizing 98,933 mitochondrial proteins. We clustered all these mitochondrial proteins available on RefSeq database in MatLab environment, using RAFTS3GROUPS algorithm. This algorithm is one implementation of RAFT3 tool, which is used to find orthologs sequences on RAFTS3 database, according to the self-score value. Here, we used a self-score 50% similarity. To infer a version of the global tree, we use all mitochondrial sequences of RefSeq database, according to neighbor-joining method. As a result, we obtained 8431 clusters, of which 2159 has two or more sequences. As expected, in the analysis we noticed that the largest clusters are the 13 most common proteins among the species, that is responsible for oxidative phosphorylation. The biggest cluster had the product COX1 protein, followed by CytB. In addition to these 13 proteins, other less common have also been identified, especially in the fungi kingdom, protist and plantae organisms. Some hypothetical proteins were also detected. The clustering relation with the inferred phylogenetic tree was consistent in most of the tree branches analyzed. It shows that the phylogenetic tree proposed can be used as a reference for future research.

Tardigrades (Ecdysozoa: Tardigrada) A general review of the record in gene database of Cytochrome Oxydase I (COI) and the Damage Suppressor gene (DSUP)

Antonio Augusto Adami Pires

Grupo de computação interdisciplinar do instituto de física de São Carlos (IFSC), Universidade de São Paulo (USP), São Carlos, SP, Brasil.

Water Bears were first described by german zoologist Johann August Ephraim Goeze in 1773, three years after this discovery, the Italian biologist Lazzaro Spallanzani appointed it as tardigrades. They are aquatics and semi-aquatics (seas and rivers) also being able to live in swamps. They are distributed in all world and there are over 1000 described species. These invertebrates are very studied in astrobiology area by NASA and ESA to be extremophiles, and they are resistant to several types of environmental selective pressures, known for their anhydrobiosis ability (survive without water) and also be resistant to temperature (-272 to 151°C), intense radiation (more than 1000Gy), incubation to organic solvents, extreme pressure (7.5GPa) and can even survive in the vacuum of space. Recently the genome *Hypsibius dujardini* (Doyere, 1840) was sequenced by two research groups: University of North Carolina and the other by University of Edinburgh. The first group suggests a large amount of exogenous genes (17.5%) probably acquired by vertical inheritance of bacteria, archea, fungi and plant due to an ability to incorporate genes in extreme situations. Changing the design of evolutionary tree for evolutionary web. But the Edinburgh group, sequencing who is available in: http://badger.bio.ed.ac.uk/H_dujardini/home/download, revealed a low percentage of exogenous genes, stating that the sequencing of the first group could be contaminated. In a current job at the University of Tokyo with tardigrade studies *Ramazzottius varieornatus* (Bertolani & Kinchin, 1993), it was also published in the GenBank the gene DSUP (Suppressor damage) responsible for radiation resistance. For this work we had used the bioinformatics software Geneious® and we did a survey of the mitochondrial gene cytochrome oxydase I (COI) that is a reference in barcoding analyses, which was identified and used the specie *Hypsibius dujardini*, genebank access KU513418 and made a comparative analysis (BLAST) in the database. The analysis revealed that the gene (COI) of this species has similarities with marine fish (98% to 99%), which is strange because there is a large taxonomic distance between the two groups. So we can infer that contamination in laboratory tests in this species can be quite frequent. The invertebrate group that has greatest similarity was a cricket *Camptonotus carolinensis* (92.03%), assuming the amount of deposits *Hypsibius dujardini* is not too high in the GenBank. The gene DSUP (Damage Suppressor) of tardigrade *Ramazzottius varieornatus* GenBank access LC050827, has only one copy which has 1338pb and translated into a protein of 334 amino acids.

Highly resolved phylogeny for Corynebacteriales

Nilson A. Da Rocha Coimbra^{1,2}, Vasco Azevedo¹, Aïda Ouangraoua²

¹ LGCM, Institute of Biological Sciences, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil; ² CoBIUS, Department of Computer Science, Universite de Sherbrooke, Sherbrooke, Quebec, Canada

The taxonomic classification of large numbers of organisms remains a hard task. In the past three decades, the 16S ribosomal RNA sequences were largely used as biological marker for phylogeny reconstruction of microbes. Nowadays, due the increase of genomic data and information produced and stored in databases, we are able to use all this whole information to reconstruct highly resolved phylogenies, by first detecting universal features in genomic data and then using them for taxonomic classification. In this work, our aim is to reconstruct the natural history and evolution of the Corynebacteriales family, in order to identify speciation in the CMNR group, the largest clade in the Actinobacteria domain. Genome data were retrieved from the RefSeq Database. The annotation in coding genes of each genome was extracted using in-house Python scripts. A customized version of Orthofinder, kindly provided by Dr. Emms, was used to cluster coding genes in families of orthologous genes. The content in coding genes of gene families was augmented using in-house Python scripts for protein prediction by homology. The phylogeny reconstruction was performed using GRIMM software and the PHYLIP software based on the augmented gene families. We collected 274 genome records of Corynebacteriales from NCBI RefSeq Database, with 1,013,515 proteins sequences in all genomes. Orthofinder predicted 27,165 clusters of homologous genes. Clusters containing paralogous genes were discarded. 22 universal, single-copy gene clusters were identified by Orthofinder. 55 additional universal clusters were obtained by augmenting existing clusters using protein prediction by homology. On total, 21,098 proteins partitioned into 77 universal clusters were predicted and used in order to reconstruct the phylogeny of Corynebacteriales.

Key amino acids in understanding evolutionary characterization of Mn/Fe-Superoxide Dismutase: A phylogenetic and structural analysis of proteins from *Corynebacterium* and hosts

Alberto Oliveira¹, Pammella Teixeira¹, Debmalya Barh², Preetam Ghosh³, Vasco Azevedo¹

¹*Universidade Federal de Minas Gerais*, ²*Institute of Integrative Omics and Applied Biotechnology*, ³*Virginia Commonwealth University*

Species from genus *Corynebacterium* can survive in a hostile environment, for example, inside a bacterial phagosome within immune system cells, such as macrophages, probably due to the production of superoxide dismutase. Recently, some studies showed this enzyme could protect bacterial cells against ROS produced by biochemical mechanisms. In addition, there are indications that in some pathogens, including some species of *Corynebacterium*, Mn/Fe-SOD may have an additional function in infection and colonization of the host. Here, we intend to conduct the coevolution analysis of amino acids from Mn/Fe-Superoxide Dismutase of *Corynebacterium* and its host, with the aim to understand the conservation and correlation among the enzymes from these organisms and consequently understand the evolutionary stages of this protein. A multiple sequence alignment of the SOD protein family (Pfam code: PF00081) was conducted by from the Pfam database and then subjected to three Itering procedures. The residue-position pairs were considered to be correlated if they passed the following thresholds: the correlation score absolute value was higher than 10 (i.e., the p-value associated with the shift in frequency is lower than 10^{-10}). The maximum likelihood method from PhyML software and multiple sequence alignment from ClustalX were used in order to understand about the phylogenetic diversion between the sequences. The comparative molecular modeling of proteins from *C. pseudotuberculosis* was performed with the software Modeller. The Validations of models were done by the PROCHECK tool that use the Ramachandran plot, the Discrete Optimized Protein Energy (DOPE) score, a statistical potential able to provide an energetic validation and by the RMSD. The tool DoGSiteScorer was used to find pockets and sub-pockets in the protein structure. Five amino acid sets were found, wherein seven residues were present in higher than 80% frequencies namely, Thr²⁴, Asn⁶⁹, Pro¹⁴⁹, Gly⁷², Gly⁷³, Met²⁵, and Gln¹⁴⁶. Two pockets were identified on the protein structure near the active site, which contain some residues observed in the sets of correlated residues. Pocket 1 is composed of Phe⁶³, Asn¹⁴⁹, Gln¹⁴⁶, Gly⁷², Asn⁷⁶ and Val¹²⁸. Pocket 2 is composed of Ile²⁴, Met²⁵ and Trp⁸¹. Analyzing the multiple alignments of Mn-SOD, it was possible to understand some divergences between bacteria and mammals in which were possible identify some key amino acids between bacterial and mammalian sequences. These amino acids were found in four loops and constitute the pocket regions that were found in our results.

Analysis of comparative genomics reveal evidence of positive selection on pathogenicity-related genes of Witches' broom disease on cocoa trees

Paulo Tokimatu¹, Marcelo F. Carazzolle¹, Paulo J. P. L. Teixeira², Daniela P. T.

Thomazella³, Leandro C. Nascimento¹, Gonçalo G. A. Pereira¹, Juliana José¹

(1) *Genomic and Expression Laboratory; Institute of Biology; UNICAMP; Brazil*, (2) *Department of Biology; University of North Carolina, USA*, (3) *Department of Plant & Microbial Biology; UC Berkeley; USA*

Financial support provided by CAPES and FAPESP.

The basidiomycete fungus *Moniliophthora perniciosa* is responsible for the Witches' broom disease, infecting cocoa trees (*Theobroma cacao*) and causing great damage on the cocoa seed industry, specially in Brazil. It is currently known 3 biotypes which can infect different hosts. The C-biotype is the most studied because it infects cocoa trees. The S-biotype infects Solanaceae plants like tomato, causing symptoms similar to the disease caused by the C-biotype. The L-biotype is found in species from the Bignoniaceae group but differently from the two other biotypes, it presents an asymptomatic infection, behaving more like an endophytic fungus. For this work we sequenced the genomes from 18 samples of *M. perniciosa* (8 samples from C-biotype, 7 from S-biotype, 2 from L-biotype and *M. roreri* as the outgroup) to investigate the evolutionary history of pathogenic genes potentially related to the difference in the host range of each biotype. After assembling, predicting and annotating our samples, comparative analysis using features like SNP counting or size of gene families showed that both nuclear and mitochondrial genomes are highly variable but have constant differences among the biotypes. It is likely that depending on their function on the infection, some pathogenic genes may suffer differential selection pressure among host species. This work focuses on the phylogenetic reconstruction, analysis of the evolutionary rate and testing for positive selection models in genes which are known to have a pathogenic role on the infection of the C-biotype on cocoa trees. We found that the genes cerato-platinin (MpCP), pathogenesis-related-1 (MpPR-1) and necrosis and ethylene-inducing proteins (MpNEP) appear in different number of copies for each biotype and that, by comparing their gene trees with a conserved tree among biotypes, are good candidates to deeper analysis. These pathogenic-related gene families showed different trees from the observed in the conserved tree, suggesting that selective processes may have act differentially on the evolution of the genes. At least one group of orthologs among biotypes showed higher divergence between L-biotype and the C- and S-samples, but a group of MpNEPs recent paralog in C-biotype and a MpPR-1 cluster shared among biotypes presented a specially higher evolutionary rate suggesting the action of positive selection. Proteins with evidences of selection might be fundamental in the process of arms-race between the host plant and the pathogen and their role can be further investigated as potential targets for the development of new technologies to contain the disease.

Analyzing molecular characteristics of small RNAs to assess the evolution of RNAi pathways

Aguiar, E.R.G.R.¹, Ferreira, F.V.^{1,2}, Olmo, R.P.¹, Vieira, K.P.O.¹, Paro, S.³, Faria, I.J.S.¹, Drumond, B, P.², Abreu, V.A.C.¹, Meignin, C.^{3,4}, Sant'anna, M.R.V.⁵, Gontijo, N. F.⁵, Moreira, L.A.⁶, Ferreira, R.S.¹, Kroon, E.G.², Imler, J.L.^{3,4,7}, Marques, J.T.¹

¹Department of Biochemistry and Immunology, Instituto de Ciências Biológicas (ICB), Universidade Federal de Minas Gerais (UFMG), Brazil, ²Department of Microbiology, ICB, UFMG, ³CNRSUPR9022, Institut de Biologie Moléculaire et Cellulaire, ⁴Faculté des Sciences de la Vie, Université de Strasbourg, ⁵Department of Parasitology, ICB, UFMG, ⁶Centro de Pesquisa René Rachou, Fundação Oswaldo Cruz (Fiocruz), ⁷Institut d'Etudes Avancées de l'Université de Strasbourg (USIAS)

A large diversity of RNA interference (RNAi) mechanisms is found in most eukaryotes. RNAi mechanisms invariably involve the formation of an effector complex known as the RNA induced silencing complex (RISC) composed of an Argonaute protein associated with small non-coding RNAs. Despite studies based on phylogeny of Argonaute proteins addressing duplication and diversification, it is unclear how these events have impacted small RNA populations. Analysis of small RNAs is challenging since the use of sequence conservation is very limited. However, small RNAs have unique molecular characteristics that depend on the pathway from which they originated. Here, we compared small RNA populations of *Aedes aegypti* (Aae), *Drosophila melanogaster* (Dme) and *Lutzomyia longipalpis* (Llo), representing three branches of dipteran insects separated by ~200 million years, to investigate the impact of Argonaute evolution in small RNA products. Since Llo lacks information about any classes of small RNAs and in Aae they are poorly annotated, we performed de novo prediction of siRNAs, piRNAs and miRNAs classes using pattern-based analyses. We identified 824 and 1,781 siRNA and piRNA clusters in Aae and 78 and 585 in Llo. Regarding miRNAs, we identified 206 miRNAs in Llo and 84 novel miRNAs in Aae. We compared the sequence and molecular characteristics of each class of small RNAs in Aae and Llo to the model organism Dme. We observed that miRNAs showed higher conservation, exhibiting similarity in sequence and molecular characteristics such as base enrichment, size profile and expression. siRNAs also presented high conservation, exhibiting similar base enrichment and size distribution. In contrast, piRNAs displayed the higher divergence. In Dme and Aae we observed U enrichment at position 1 of antisense and A enrichment at position 10 of sense small RNAs, which were not observed in Llo. Furthermore, we also noticed significant differences in the size distribution of piRNAs comparing the three insects. However, we observed 10-nt overlap between 5' end of reads in opposite strands, showed to be the most conserved piRNA feature among dipteran. Our results suggest that while miRNA and siRNA pathways are highly conserved displaying similar molecular features, piRNA pathway is the most divergent, showing discrepancy in size and base enrichment. These results showed correlation with divergence and expansion of Argonaute proteins in insects analyzed. Thus, studying small RNAs can directly contribute to evaluation of RNAi pathways regarding the impact of protein changes in small RNA products.

Financial Support: CAPES, CNPq and FAPEMIG

Evaluation of the accuracy of the ML-relate kinship analysis program when no parents are sampled

Letícia F. Lima, Renata Schama

Laboratório de Biologia Computacional e Sistemas, Oswaldo Cruz Institute- Fiocruz

Knowledge of the genealogical relationships among different family groups can provide important information for the study of evolution and behavioral ecology. In parentage analysis of natural populations, numerous methods have been developed for inferring the relationship among individuals aiding conservation management in zoos and wildlife reserves. In most cases the genotypes of both offspring and potential parents are present in the sample and this data makes it much easier to assign the offspring to parents. This kind of method may also be used to better understand oviposition and migration behavior in insects. Information such as these can be of assistance for new vector control strategies. Nevertheless, in insects, most of the time the collection of samples does not involve collecting parent and offspring together; which makes the inference of kinship relationship less accurate. Also, polygamy, another characteristic of many insects, increases the difficulty of the analysis. The presence of half siblings in the sample makes it difficult to understand the family groups. In this study we use simulation data created using real data parameters (number of loci, alleles and genetic diversity) to test the accuracy of the program ML-relate to detect full siblings. The ML-relate program is widely used in parentage analysis and uses a maximum likelihood framework to find the most likely relationship between two individuals. To simulate the data, with the COLONY program, we used the genetic structure and diversity of the mosquito *Aedes aegypti* population from Rio de Janeiro. We simulated data containing different percentages of half-siblings (10%, 20%, 30%, 40% and 50%) to assess how the different scenarios would affect the program's classification of kinship relationships. We evaluated the error rates and accuracy of the program under the different scenarios. We found type I error to be low in all scenarios (around 0.021%). Type II errors fluctuated and were much higher (around 0.51%), indicating that for natural data a large number of full siblings would not be categorized as such. Nevertheless, accuracy was 76%, well within the range found for other programs and organisms analyzed. Since type I error was low and type II high we conclude that although the estimation of family groups will be smaller than the real ones we can be sure of the full sibling relationships estimated by the program and therefore of the inferred oviposition behavior. Nevertheless, these results will impact conclusions regarding the extent of migration by females, underestimating it.

Assortative Mating in Brazilian Populations

Isabela Alvim, Hanaisa de Plá, Eduardo Martin Tarazona

Human Genetic Diversity lab, UFMG

The admixed Brazilian population shows a trend of marriages between individuals with similar socioeconomic status, educational level and genomic ancestry. This behavior violates the assumption of random marriages of several statistical models in genetics populations and medical genetics, which can lead to misleading results in quantitative genetics analysis and demographic inferences. Therefore, the knowledge about the trends of marriages in Brasil is crucial to the progress of Brazilian population genomic studies. In previous studies our group verified the occurrence of ancestry-assortment for the three Brazilian populations of the EPIGEN-Brasil (The Latin American initiative in population genomics and genetic epidemiology): Salvador – BA, Pelotas – RS, Bambuí – MG. In this project we investigated the patterns of ancestry-assortment for different socioeconomic status and educational levels in these populations. The 6.487 samples from the three population-based cohorts were genotyped for 2.5 million SNPs (single nucleotide polymorphism) and each cohort was stratified in 3 categories based on individual educational level. Homozygosity excess in SNPs that are highly differentiated between ancestral populations indicate ancestry-assortment, therefore we estimated the Spearman's rank correlation (ρ) for the ancestry informativeness and the homozygosity excess estimated for each SNP. Our study finds that the ancestry-assortment is significantly affected by educational level. Salvador and Pelotas showed a crescent pattern with the most remarkable correlation in the high educational level ($\rho = 0.07$, $p\text{-value} < 2.2e-16$; $\rho = 0.34$, $p\text{-value} < 2.2e-16$ respectively) compared to the low ($\rho = 0.01$, $p\text{-value} < 2.2e-16$; $\rho = 0.11$, $p\text{-value} < 2.2e-16$) and middle levels ($\rho = 0.05$, $p\text{-value} < 2.2e-16$; $\rho = 0.23$, $p\text{-value} < 2.2e-16$). The opposite is seen in Bambuí, that shows a more expressive evidence of ancestry-assortment at the lower educational level ($\rho = 0.14$, $p\text{-value} < 2.2e-16$) compared to the middle and high levels ($\rho = 0.05$, $p\text{-value} < 2.2e-16$; $\rho = 0.03$, $p\text{-value} < 2.2e-16$). Thus, our results enables that the adjustments of mathematical models for Brazilian population genomic studies consider those regional differences in the assortative mating.

Insights into the population history of free-living bacteria as counted by their CRISPR inventory

Julliane Dutra Medeiros^{1,2}, Laura Rabelo Leite^{1,2}, Francislon Silva^{1,2}, Victor Satler Pylro², Gabriel Rocha Fernandes², Guilherme Oliveira³, Sara Cuadros-Orellana²

¹ Universidade Federal de Minas Gerais, UFMG, Minas Gerais, ² Centro de Pesquisa René Rachou, Fiocruz, Minas Gerais, ³ Instituto Tecnológico Vale- Desenvolvimento Sustentável, Pará.

The CRISPR-cas is an adaptive and heritable immune system encoded by prokaryotes, which provides insights into genetic diversity within bacterial populations. The CRISPR array contains short repeats interspersed with unique spacers. The spacers are small pieces of DNA derived from foreign nucleic acid and represent chronological records of viruses infecting the cell. We investigated the population history of *Hydrotalea* sp. based on their CRISPR repertoire. Single-cells from an early-stage acid mine drainage were sorted and amplified following the Bigelow SCGC pipeline. Six single-amplified genomes (SAGs) of Chitinophagaceae were partially reconstructed. The phylogeny and average nucleotide identity analysis suggest that all SAGs represent a same species within the genus *Hydrotalea*, at the species boundary of *H. flava*. They form two clades representing different strains and cells inside the same clade are clones. CRISPR-cas were recovered across SAGs. Cells L11, B16, P17 and J04 encode two co-existing CRISPR loci, type I-B and II-C. Newly described repeat sequences are shared by SAGs, and all spacer sequences are divergent among the cells from two clades. Considering type I-B, there is evidence of loss of older spacers in SAG B16. Analyzing the genome content of other Chitinophagaceae species we observe only type II CRISPR loci, our hypothesis is that the SAGs or their ancestor cells had a CRISPR-Cas type II-C that is conserved in the Chitinophagaceae family, and also a CRISPR-Cas type I-B system that was likely acquired through HGT. It is important to mention that the SAGs harbor transposases and features related to conjugative transposons (CTns). The CRISPR based heterogeneity of population documented here, as well as the loss of spacers, suggest that organisms are adjusting to a quickly changing selective pressure in a microhabitat scale. A very likely form of such selective pressure is phage predation.

Financial Support: Vale

Evolutionary origin of the proteins involved in the entry of and defense to the Ebola virus in the host cell

Lopes, E. N.¹, Sakamoto, T.¹, Ortega, J.M.¹

¹ Laboratório de Biodados, Departamento de Bioquímica e de Imunologia, ICB;
Universidade Federal de Minas Gerais (UFMG).

The Ebola virus (EBOV) belongs to Filoviridae family and is the causative of a devastating disease, with a mortality rate of about 50-90%, known as Ebola hemorrhagic fever. The first symptoms developed by infected patient are fever, malaise and muscle pain, and can be followed by bleeding and organ failure. The known virus hosts are humans as well as non-human hosts, like other primates and fruit bats. The Ebola mechanism of infection is complex, which makes it versatile. In this work, we revised the scientific literature that unravels the infection mechanism of EBOV and collected the host proteins known to mediate infection by their interaction with the viral proteins. We also analyzed the homologous of each collected human proteins along the taxonomic tree using the software SeedServer and Genesis to infer their clade/epoch of origin. Several researches indicate the relevance of Ebola glycoproteins GP1 and GP2 as targets for evolutionary researches and therapeutics, since they participate in virus infection mechanism on host cells and they are conserved throughout EBOV strains. The glycoproteins GP1 and GP2 interact with some main proteins on host cell: CLEC4M, CD209, EGFR, NPC1, NPC2, CLEC10A, TYRO3, FOLR1, CTSB, CTSL, which participate directly or indirectly of the infection mechanism. Conversely, the defense capability of the host to the Ebola virus is much debilitated because of the genomic diversity of the five known EBOV strains and the virus interaction with the host immune system proteins responsible for signaling an invader agent. They are: MHCI, EGFR, TETHERIN, ITGA1, ITGA2, ITGA3, ITGA4, ITGA5, ITGA6, ITGAV, ITGB1, ITGB3 and ADAM17. This interaction of the virus with proteins causes a delay leading to low expression of the immune response. The result of the evolutionary origin analysis of these proteins showed that the virus could infect even vertebrates, suggesting that animals such as fish and amphibians could be infected and retransmit the virus to other hosts such as man. Regarding the immune response, the evolutionary analysis suggests that it began to be developed in Eukaryota but the full response was achieved more recently, only in Eutheria (placental animals like dog, elephant, primates, among others). The origin and the Ebola virus cycle in hosts is still uncertain, therefore these results bring questions about new reservoirs and suggest that defense against it is more recent than the possibility of infection.

Financial Support: CAPES, FAPEMIG

Study of new molecular markers for Phylogenetic reconstruction of the black fungus in humans

Edgar Lacerda de Aguiar¹, Cláudia Barbosa Assunção², Rachel Basques Caligorne²

¹ Bioinformatics and Systems Laboratory, Federal University of Minas Gerais, Brazil

² Mycology Lab, Research institute of the Santa Casa de Belo Horizonte, Brazil

Melanized hyphomycetes, or black fungi, are the etiological agents of opportunistic infections diseases, grouped into three classes: chromoblastomycosis, mycetoma and phaeohyphomycosis. Recently PCR methods (Polymerase Chain Reaction) have been successfully used for fungi identification, including melanized fungi of medical importance. Phylogenetic inferences, or taxonomic, organize knowledge on biological diversity, from the relationships among groups and knowledge of evolution of morphological, behavioral, physiological, cytogenetic, and molecular organisms. For the realization of phylogenetic analysis is necessary to choose a method to use, there are several statistical methods to carry out the phylogenetic analyzes each according to need analysis, specificity, and targets. Addition to the method is necessary to select a molecular marker to be used in phylogenetic analyzes. In fungi the ITS1 and ITS2 domains (Internal transcribed space) of the small subunit of the ribosome (18S-rDNA) can be well used in comparative analysis between species, as these regions are transcribed, but not translated, which allows these domains accumulate mutations, turning highly variable regions. Some authors have characterized two kinases of *Paracoccidioides brasiliensis*, responsible for inhibition of the Translation Initiation factor eIF2, EIF2AK1 and EIF2AK2, in response to cellular stress, such as temperature change, shortage of amino acids and stress Osmotic. Given the importance of these proteins, this research project aims to compare sequences of the kinases in different fungal classes, in order to design molecular markers for identification of species and even to aid in diagnosis of diseases caused by fungi. Through the analysis of sequences deposited in GenBank was observed the occurrence of genes kinases EIF2AK1 and EIF2AK2 in black fungi that are human pathogens. The result of the preliminary phylogenetic analysis confirms the results seen by Caligorne and collaborators, in which *Fonsecaea* genus are part of a phylogenetic arm in relation to other genera, *Cladophialophora*, *Phialophora*, *Rinocladiella*. The results that can come with discovered new markers can help bring new information related to the classification, to demonstrate the diversity of etiologic agents of chromoblastomycosis, phaeohyphomycosis and mycetoma.

Molecular Docking and Structural Optimization of Bioactive Compounds from Natural Products Against UAP-1 of *L. brasiliensis*

Sabrina Silva Mendonça, João Marcos Galúcio, Kelly Christina Ferreira Castro,

Kauê Santana da Costa

Federal of University of western of Pará

Leishmaniasis are classified by the World Health Organization (WHO) as a neglected tropical diseases and receive low research investments for developing new methods for treatment and prophylaxis. Molecules derived from natural products are an interesting source for drug synthesis and design, but we little know about their macromolecular targets. Thus, computational methods have high importance to study and analyze the interactions of drug-like molecules, which are candidate to become prototype of new drugs. Assuming that structurally similar molecules exhibit similar biological activity thus, we searched in different public databases for analogs to the essential oil molecules obtained from *Piper marginatum* which showed good inhibitory activity against *Leishmania* promastigote cultures. These compounds were extracted of leaves from *Piper marginatum* by hydrodistillation and then identified by gas chromatography coupled mass spectrometry. With the molecular structure of the compounds, we then performed structural alignment in Marvin Sketch seeking desired values of Tanimoto similarity. Thereby, we found an interesting structural similarity with inhibitor of N-acetylglucosamine pyrophosphorylase (UAP-1) of *Trypanosoma brucei* with the molecule 3,4-methylenedioxypyropiophenone, which constitute 21% of essential oil. Searching in Blastp, we also found 41% of identity between *T. brucei* protein with the homologous UAP-1 from *Leishmania brasiliensis*. Then, we modeled the structure of LbUAP-1 by homology in Modeller, using as reference the homologous protein of *T. brucei* (PDB ID 4BQH). The model obtained was further evaluated by Ramachandran plot in PROCHECK program, Verify3D, and also by the atomic non-local energy profile in the ANOLEA plot. Analyzing the two cavities, we noted that both shares similar topography and electrostatic potential map. Then, molecular docking simulations of the compound 3,4-methylenedioxypyropiophenone, against the modeled UAP-1 protein of *L. brasiliensis* was performed in Molgro program using the MolDock Optimizer algorithm. The cavity of protein was pre-selected in the LbUAP-1 using as reference the spatial coordinates of homolog structure in *T. brucei* and it showed similar bidding mode with the crystal inhibitor. Therefore, we designed new inhibitors using the structure with optimized affinity to the bidding site and the interactions were analyzed by Biovia Discovery Studio program. This compound from *Piper* essential oil could be used as fragment-based optimization and synthesis of new inhibitors against *Leishmania* species.

Cluster analysis of AcrB protein molecular dynamics conformations

Núbia Souza Prates¹, Adriano V. Werhli², Karina dos Santos Machado²

¹*Biological Sciences Institute - Universidade Federal de Minas Gerais*

²*Center of Computacional Sciences - Universidade Federal do Rio Grande*

Molecular dynamics (MD) simulation is a computational method widely used to study biological macromolecules at an atomic level. The classical MD simulation is based on numerical solution of Newton's equations of motion. Typically this method may generate as result a large amount of data to be analyzed. These data consists in thousand of conformations that can be used, for instance, to understand the macromolecule flexibility and/or to incorporate this flexibility in docking studies. Due to the complexity of these data, cluster analysis is one approach that can be applied to MD trajectories to reduce this data, identifying representatives conformations of each cluster. First we performed a 50 ns MD simulation of AcrB efflux pump protein using GROMACS package. Protein conformations were taken from the 50 ns MD trajectory with intervals of 50, 20 and 10 ps totaling 3 datasets of conformations having 1,000, 2,500 and 5,000 protein structures each. Thus, in this work we propose to compare three different clustering algorithms for MD simulations considering as input 3 datasets of protein conformations. We applied the clustering algorithms Gromos, Single-Linkage and Jarvis-Patrick implemented in GROMACS. Using the Root-mean-squared deviation (RMSD) as similarity measure with a cutoff from 0.1 to 0.5 nm. Increasing the RMSD values we notice that for Single Linkage and Jarvis-Patrick the number of clusters decreases and all structures tends to remain in a unique cluster. The same was observed for Gromos, where higher values of RMSD implies in all the conformations in the same cluster or leading to the formation of fewer overpopulated clusters. The amount of clusters composed by a single structures was higher in Jarvis-Patrick clustering results compared with the others algorithms using the cutoff 0.17 nm. In the implementation of the clustering algorithm in GROMACS there is no internal or external cluster validation indices to evaluate which algorithm presented the best clustering results. Thus, we consider as best clustering results those clusters with more than 2 structures, fewer singletons clusters or those clusters with approximately the same distribution of protein conformations. Compared to others algorithms Gromos produces significant clustering results, presenting the same behavior with all datasets. We also evaluate the Free Energy of Binding (FEB) by molecular docking experiments considering as receptor representative conformation of each cluster and a ligand called NLM02. Additionally, Gromos presented the best results of docking experiments with a minimum FEB of -9,8 kcal/mol, while -9,4 for Single-Linkage and Jarvis-Patrick.

Supported by: CAPES, CNPq.

Ligand-Based Pharmacophore Modeling and Virtual Screening of Plant-Derived Ligands for the Alpha-Amylase and Alpha-Glycosidase

Heitor Cappato Guerra Silva, Nilson Nicolau Junior, Foued Salmen Espindola

Institute of Genetics and Biochemistry, Federal University of Uberlândia. Uberlândia, Minas Gerais, Brazil

Natural antioxidants compounds have been associated with reduction of postprandial hyperglycemia by blocking enzymes involved in the carbohydrates digestion, such as alpha-amylase and alpha-glycosidase. Furthermore, preventing or delaying the absorption of glucose by inhibiting glycoside hydrolases in the digestive organs may represent a promising approach in the treatment of diabetes and its complications. Thus, the aim of this work was search for new plant-derived compounds with pharmacological potential to inhibit this glycoside hydrolases based on the pharmacophore model. The pharmacophore modeling was performed with the aid of vROCS 3.2.0.4, this model contains information about shape and chemical properties extracted from the fluconazole molecule. The ligand library used in this research are originated from ZINC database, that have been carefully selected a natural compounds subset, totaling 180.303 compounds. In order to perform the virtual screening, the ligand library was prepared with the OMEGA 2.5.1.4, which was used to generate conformer libraries. Pharmacophore model validation and virtual screening of the conformer libraries were performing using vROCS. The pharmacophore model was previously validated using the ROC (receiver operating characteristic) curve and AUC (area under the curve). The AUC extract from the ROC curve graph it is simply the probability that randomly chosen bioactive compounds have a score higher than randomly chosen inactive compounds. In order to generate the ROC curve and the AUC value, biologically active ligands against alpha-amylase (PDB id: 1SMD) and alpha-glycosidase (PDB id: 1OBB) were obtained from ZINC database, and the decoys were generated on the DUD-E online platform. After validation, the conformer library previously generated was submitted to the pharmacophore model and the top 500 ligands of each, based on the TanimotoCombo score, were selected. The best-scored ligands will be used to perform a molecular docking against human alpha-amylase and alpha-glycosidase as the next step of this research.

Financial Support : Fapemig e CNPq.

Structural Analysis of Alba Proteins of *Leishmania infantum*

Elvis Santos Leonardo¹, João Marcos Galúcio, Élcio Souza Leal, Kauê Santana da Costa¹ Jerônimo Lameira²

¹Universidade Federal do Oeste do Pará – UFOPA ²Universidade Federal do Pará - UFPA

The Alba are a superfamily of proteins involved in binding to the RNA/DNA that share a common domain. We modeled the structure of Alba13 and Alba20 proteins using the homology and threading approaches, respectively. In *Leishmania* species, the Alba20 and Alba13 are constitutively expressed in amastigote and promastigote forms of *Leishmania* life cycle, and previous studies showed that it is involved in the regulation of expression of δ-amastin gene. So considering it show to be interesting target for rational drug design against these parasites, we performed multiple structural analyses in these proteins, using protein prediction, mutational analysis and protein-protein docking. The amino acid sequences of Alba 13 and Alba20 were obtained in the TriTrypDB databank. We used the Modeller to predict the Alba13 structure using as template a hypothetical protein of *Arabidopsis thaliana* (PDB ID: 1VM0, chain A) which conserved the Alba domain and showed 36% of identity in the sequence alignment. The C-terminal region without homology was obtained by threading in I-TASSER server. The Alba20 structure was modeled only by *threading*, and all structures were refined in 3000 cycles of conjugated gradient algorithm and then validated by Ramachandram plot and ANOLEA energy profile. We docked the Alba20 and Alba13 in Rosie server. In order to analyze the stability effect of mutations in the structures, we also performed an alanine scanning in both proteins using FoldX. Our model showed a good energetic profile and a satisfactory stereochemical quality, with number residues in the favorable regions of Ramachadran plot. The residues of Alba domains interact in both proteins, stabilizing the interaction in the dimer complex. Our mutational analysis showed that most mutations in the C-termini region showed to be high destabilizing in Alba20 structure, and in that region is located the RGG box motif which interacts with the 3'UTR region of δ-amastin transcript.

Key-words: Alba, Prtozoa, Modeling

Non-Homology-Based Prediction of Protein Target Regions by Logistic Regression

Gustavo Santos de Oliveira¹, Marcos Augusto dos Santos¹, Vasco Ariston de Carvalho Azevedo¹

¹*Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte – MG, Brazil.*

Influenza A is a RNA virus responsible for multiple types of health problems in different hosts, such as chickens, pigs and humans. The two proteins involved in its pathogenicity are hemagglutinin and neuraminidase. These are proteins, which triggers the virus' entrance inside the cell. Treatment options has been a challenge due the high evolutionary rate of the virus proteins. The prediction of potential target sites of key virus proteins has been a difficult task due lack of homology between sequences. In this sense, here it is proposed a new methodology to identify conserved residues associated with viruses' proteins by means of logistic regression. Due to its known activity related to the virus internalization, hemagglutinin protein sequences of Influenza A H5N1 and H3N2 were selected as a positive highly pathogenic and less pathogenic model, respectively. From UNIPROT, 5466 sequences for H5N1 and 259 for H2N3 were obtained. The model was built using MATLAB®, in an aproach where a sliding window was designed to count all possible triplet residues for each sequence. The matrix of all possible triplets for all sequences were submitted to an initial Singular Value Decomposition (SVD) for sampling homogeneization, and logistic regression, aiming to find triplets associated with the enhanced pathogenicity effect for H5N1 hemagglutinin compared to H2N3. The present aproach was able to detect critical regions associated with the high pathogenic hemagglutinin activity. That is the case of RRKKR, which is located within a connector loop between HA1 and HA2 subunits and is known as a target for proteolytic cleavage responsible for hemagglutinin activation. Hemagglutinin allows virus internalization thanks to its ability to bind the host membrane and its capacity to bend itself in low pH, permitting the virus to get inside the cell. The methodology detected, also, the triplet SII, located within the alpha-helix of HA2 subunit, in a region that acts as a hinge that tights the interaction between the two subunits after conformational change due pH decrease. Other sections detected were SNEQG, a terminal HA2 region, responsible for the formation of host membrane pores, and KIA, located in a beta-sheet region in HA1 and associated with the protein stabilization in acidic environments. In conclusion, this methodology was able to predict key regions for the hemagglutinin mechanisms of action and to corroborate with some authors that highlight the entrance mechanisms as more significant than the host cells recognition mechanisms by hemagglutinin, for the virus pathogenicity.

Selecting structure-based virtual screening hits using chemoinformatics tools: a case study with HIV-1 reverse transcriptase

LH Santos¹, RS Ferreira², ER Caffarena¹

¹Fundação Oswaldo Cruz, Programa de Computação Científica, Rio de Janeiro, RJ, Brasil

²Universidade Federal de Minas Gerais, Instituto de Ciências Biológicas, Belo Horizonte, MG, Brasil

Reverse transcriptase (RT) is a major drug target for the treatment of HIV infection. RT is inhibited by two inhibitor classes: the nucleoside RT inhibitors (NRTI) and the non-nucleoside RT inhibitors (NNRTI). The NRTIs, when bound to RT active site, obstruct the conversion of single-stranded RNA to double-stranded DNA provirus, which is then integrated into the human genome. Whereas, NNRTIs, bind to an allosteric site, causing conformation changes that impair DNA synthesis. Despite the large number of drugs targeting HIV-1 RT, problems like resistance, toxicity, and especially mutations turn the development of more effective and less toxic inhibitors an urgent matter. Due to the availability of several crystal structures, structure-based virtual screening might be a preliminary effort to rational drug design of RT inhibitors. However, after the screening of a compound library through molecular docking, there is still the task of analyzing and interpreting hundreds to thousands of ranked compounds. To this task, chemoinformatics tools use unique representations of chemical structures in the form of descriptors, utilizes metrics of similarity, and apply statistical and other techniques to establish relationships between chemical structures and their properties. Therefore, this study aimed to identify a subset of hit compounds after applying chemoinformatics tools to the virtual screening outcomes of an extensive compound library using the non-nucleoside binding pocket (NNBP). The lead-like now ligand subset from the ZINC database was chosen as the screening library, at the time composed of 2,797,315 compounds. All compounds were submitted to molecular docking using DOCK6.6. The use of molecular docking method ensured that the subsequently chosen compounds, would fit, and possibly have interactions with the NNBP. The screening was performed using an RT structure (PDB: 4G1Q) bound to the known NNRTI, rilpivirine. This structure achieved the best performance amongst ten others in the prior assessment of DOCK6.6. Among the 50% top-scoring compounds (GridScore of -40.52 kcal/mol or lower), we chose two samples of 5,000 compounds each. One sample containing compounds ranked by the lowest (more favorable) GridScore, and another containing randomly selected compounds to explore the possibility of finding out suitable candidates in the remaining database. With the help of the R package ChemmineR, clustering of the sets was done to identify discrete similarity groups using the binning clustering function with Tanimoto coefficient at a cut-off of 0.55. In the end, 68 and 86 chemical diverse compounds were obtained from each subset, respectively. To further filter the remaining compounds, we selected only compounds with one or more specific interactions with important NNBP amino acids, known to interact with known inhibitors. Therefore, we were able to propose a more manageable number of hit candidates for further testing from the large initial screening library, making use of valuable metrics to ensure the chemical diversity of ranked compounds by docking.

Supported by: Fiocruz, Capes.

Characterization of the EF-IV (LepA) influence on mRNA translation by in-silico cell-free protein expression

Anton Semenchenko, Bárbara Zanandreiz Siqueira de Mattos, Guilherme Oliveira,

A. P. F. Atman

Aplasys, Centro Universitário Newton Paiva, Vale Technology Institute, CEFET-MG

The LepA is a high conserved protein and an essential elongation factor for ribosome function. Its dominant role is to facilitate the back-translocation of the ribosome when defective translocation occurs during mRNA translation. We demonstrate the results of the in-silico characterization of the error correction role of EF-4 as well as its catalytic and toxic effects in virtual cell-free protein expression systems. The EF-4 role in ribosome back-translocation is described using computer simulation and validated by the experimental data. The results indicate that catalytic effects of the EF-4 originate outside of the translocation mechanism and must be introduced separately from the translocation model. The error correction and toxic effects of this elongation factor clearly reproduce the experimental evidence and confirm the back-translocation origin of this effect. These results together with the calibrated simulation allow for quantification of the catalytic and toxic effects of EF-4. In order to study the dynamics of the protein synthesis under influence of the elongation factor IV (LepA) we employed calibrated and validated computational approach to cell-free protein expression systems. The calibration procedure is demonstrated by the simulation of the in-vitro 5 hours of Luciferase production within the environment equivalent to Rapid Translation System RTS 100 by Roche. In addition to calibration, the model is validated by the simulation of the Edeine antibiotic effect. The integration of the hybrid model of mRNA translation with the model of ribosome translocation is presented. The modeling technique employs the combination of the 3D cellular automata and agent-based simulation. Stochastic nature of the biochemical processes orchestrated by the ribosome are represented by the set of Markov chain that reflect the distinct treatment of the cognate, near- and non-cognate tRNA by the ribosome. The flexibility and adaptability of the presented model combined with computer simulation is illustrated by the ability to reproduce a number of behaviors observed in the in-vitro experimentation. The application of the resulting computational system is illustrated by the virtualization of the cell-free protein expression kits.

GAPC1 and HSP90.1 proteins may have an important role in the oxidative stress in *Saccharum* spp.

Felipe de Lima Almeida¹, Kellya Francisca Mendonça Barreto², João Paulo Matos Santos Lima³, Katia Castanho Scortecci⁴

Instituto Metrópole Digital (IMD)^{1,3}, Universidade Federal do Rio Grande do Norte

Abiotic and biotic stresses have a huge impact in sugarcane plant growth as its affects yield production. These stress conditions change plant development. Moreover, one of its consequences is the increase of H₂O₂ production as well as other reactive oxygen species (ROS). In order to understand better this oxidative stress, then sugarcane plants were grown in H₂O₂ for 8 hours. After that, roots and leaves were isolated for proteomic analysis. The data obtained showed a differential protein expression for Glyceraldehyde-3-phosphate dehydrogenase (GAPC1) and Heat shock protein 90.1 (HSP90.1) in the oxidative stress condition. Due to this, these two proteins were chosen for further analysis using bioinformatics tools in this work. The results obtained allowed us to observed that these two proteins GAPC1 and HSP90.1 make an interactome. Furthermore, it was observed five proteins with intersection in this interactome: SCE1, GRF1, Q38942, RAD23A and UBQ3. This data allowed us to propose a following model for action: the H₂O₂ may inactive the GAPC1 protein, which promotes its oxidation and induce the PLD δ activity, then the phosphatic acid is accumulated in the cell. When GAPC1 is reduced, the glyceraldehyde-3-phosphate is converted to 1,3-bisphosphoglycerate and NADH is produced. Then, GAPC1 oxidation/reduction are a result from stress condition, which may induces SCE1 protein to modify the signal transduction in cytosol, consequently UBQ3 protein may interacts to RAD23A in order to correct possible lesion in DNA due to oxidative stress. Besides, the nucleoporin Q38943 may have a role as exportation factor for mRNAs promoting GRF1 to regulated DNA expression in response to abiotic stress. Moreover, the role for HSP90.1 may be related to keep protein stable as well as preparing plant to be able to tolerate this stress condition. In a nutshell, this model showed how these proteins may act in the cells in order to reduce the negative impact from oxidative stress in cell, which will be important in plant metabolism to tolerate these adverse conditions.

Financial support: CNPq, CAPES.

Structural pattern detection for engineering more efficient enzymes for second-generation biofuel production

DCB Mariano¹, TS Correia¹, JRPM Barroso¹, RC de Melo-Minardi¹

¹ LBS - *Laboratory of Bioinformatics and Systems. Department of Computer Science. Federal University of Minas Gerais. Belo Horizonte, Brazil*

β -glucosidase (E.C. 3.2.1.21) is the main enzyme in the process of the second-generation biofuel production. It acts synergically with endoglucanases and exoglucanases to degrade cellulose in glucose, which will be used for bioethanol production. β -glucosidase plays a key role in the last step of this enzymatic system, converting cellobiose, a disaccharide that inhibits both endo- and exoglucanases, in glucose. However, most of the β -glucosidases known are inhibited by glucose. Recently, our group has preceded a systematic literature review (SLR) to evaluate the state-of-art of β -glucosidase researches. In this SLR, we collected 23 sequences and three-dimensional structures of β -glucosidases with high tolerance to glucose inhibition. In this work, we propose an analysis of these collected structures to detect patterns that can be used to engineering of β -glucosidases with high catalytic efficiency, and also, to detect possible glucose-tolerant β -glucosidases in data obtained by high-throughput platforms of sequencing. We collected 3,991 β -glucosidases sequences of the GH1 family (described in the literature as more efficient for biofuel production) from UniProt and performed homology modeling. In the first step, we analyzed patterns in the primary structure using amino acid k-mer frequency and singular value decomposition (SVD). However, the sequences were not sufficient to cluster the glucose-tolerant β -glucosidases. To investigate whether the patterns appear in other structural levels, we performed structural alignment of 21 three-dimensional GH1 β -glucosidases structures with the β -glucosidase from termite *Neotermes koshunensis* in complex with cellobiose (PDB: 3VIK). Then, we collected the amino acids at a distance of 6 Å, 6.5 Å, 7 Å, 7.5 Å, and 8 Å of the substrate, and submitted them to the software aCSM-ALL to detect contact frequencies based on cutoff of atomic distances and on physicochemical information. We reduced the noise of the atomic distances collected with SVD, and with this information we constructed a glucose tolerance signature to identify high efficient β -glucosidases for biofuel production. We also characterized an ideal active site based on multiple alignments of high tolerant β -glucosidases. The glucose tolerance signature can be used to detect proteins potential targets for cellulose degradation. Also the characterization of an ideal active site based on glucose-tolerant β -glucosidase data can be useful for enzyme engineering with high catalytic efficiency and may help shed light on the second-generation biofuel production.

Supported by: FAPEMIG, CNPq, and CAPES (51/2013 - 23038.004007/2014-82).

Using sequence weighting to improve residue correlation analysys

Lucas Carrijo, Lucas Bleicher

Institute of Biological Sciences, Federal University of Minas Gerais, Brazil.

Analysing a multiple sequence alignment at the residue level, apart from the conserved positions, there are other patterns that are also indicative of functional importance and reflect functional divergence within a homologous protein family due to gene duplication. In families that have subfamilies with distinct functional specificities, some positions can be conserved only in a particular subfamily, or the conserved amino acid can be different for each of the subfamilies. This suggests that the role of this residue relates not to the global function of the family, but to functional specificities of that group. In these cases, it is reasonable that such specificities are not determined by the presence of a single residue, but by a group of residues, and this group will emerge from residue correlation analysis since a sufficient amount of proteins show the same specificities. However, some protein families have subfamilies less represented in terms of amount of sequences in the alignments. Meantime, these alignments used to come full of redundant sequences, many times mutants or variants of the same sequence, originally mainly from model organisms. This redundancy in the alignments tends to introduce bias to analysis with a statistical mean like the correlation methods. In this way, the present work has as objective to compare the effects of distinct approaches aiming the decreasing of redundancy in multiple sequence alignments: sequence weighting and filtering by maximum identity. Besides, this work also proposes approaches to make the correlation calculations compatible with sequence weighting, in order to improve analysis of residue conservation and correlation. Sequence weighting was capable of highlighting frequencies of amino acids specific of less sampled subfamilies, while decreasing the frequencies of amino acids present in redundant sequences. The adapted calculations were capable of detecting such differences, providing a good alternative to conservation and correlation analysis in alignments that are less representative of the actual protein diversity existent in nature.

Mutational Analysis of the Virion Infectivity Factor (Vif) of HIV-1 subtype F2 and its influence on the interactions with APOBECs

João Marcos Galúcio¹, Elcio Souza Leal², Sabrina Silva Mendonça¹, Kauê Santana da Costa¹

¹Universidade Federal do Oeste do Pará; ²Universidade Federal do Pará.

The Virion Infectivity Factor is a protein of HIV extremely important for viral infectivity and replication. The induction of the proteasomal degradation of antiretroviral proteins APOBECs is the most important function. The APOBEC3 family comprises cytidine deaminases which are differentially expressed in HIV-1 susceptible cells. The Vif alleles neutralize A3G and A3F efficiently, but display differences with respect to the inhibition of A3H. Two subtype F Vif variants show the highest activity against A3H; its recognition requires the residues F39 and H48 inVif structure. Alterations on this residues show to be crucial to the infectivity difference of subtype F2 against A3H protein. This work investigate possible changes in the structural stability and in the electrostatic potential of native Vif and relates it to changes in binding with A3H, A3G and A3F experimentally verified. The Vif of HIV-1 subtype F2 was modeled by homology in the Modeller software and we generated F39S, F39S/H48N and F39V/H48N variants using Fold X, to compare the free energy and electrostatics potential maps between native and mutant forms. The biologic impact of the mutations was predicted in Provean server. Then, we performed a Alanine Scanning in the motifs involved on the interaction with A3G and A3F to verify their contribution in the stability of Vif. The results showed that F39S mutant don't generate changes in the electrostatic potential and structural stability of the protein. However, the mutations F39S/H48N and F39V/H48N were classified as stabilizing ($\Delta\Delta G = +1.60$ kcal and $+1.56$ kcal/mol, respectively). We observed inversions and other alterations in the mutants electrostatic potential maps that can influence directly the interactions with A3H protein. These two mutations induced the formation of four new hydrogen bonds between the residues R41 and E45, H42 and S46; and, V51 and H43. However, the Provean Score indicated these variants could be deleterious for Vif function. In the alanine scanning, the motifs 40YRHHY44 and 161PPLP164 (A3G binding), 14DRMR17 (A3F binding) and 21WNSLVK26 and 55VHIPLKDDSL64 (A3G/A3F binding) show to be highly important for structural stability of Vif, mainly 55VHIPLKDDSL64 ($\Delta\Delta G = +15.61$ kcal/mol) and 161PPLP164 motifs ($\Delta\Delta G = 9.17$ kcal/mol). According to experimentally verified, residues 39F and 48H are very important for A3H recognition and changes in their positions can alter electrostatic properties and structural stability of Vif in subtype F2.

Molecular interactions between NF-κB and thiopheneacetamide during mycobacteria infection

VS Silva¹, FM Vergara^{2,3}, MG Henriques^{2,3}, ER Ca arena¹

1 Computational Biophysics and Molecular Modelling Group, PROCC, Fiocruz;

2 Institute of Drug Technology (Farmanguinhos), Oswaldo Cruz Foundation (FIOCRUZ), Rio de Janeiro, Brazil; 3 National Institute for Science and Technology on Innovation on Neglected Diseases (INCT/IDN), Center for Technological Development in Health (CDTS), Oswaldo Cruz Foundation (FIOCRUZ), Rio de Janeiro, Brazil

The nuclear factor kappa B (NF-κB) pathway is a key role on the host response against many pathogens, as *Mycobacterium tuberculosis*, the etiological agent of tuberculosis (TB). TB is one of the oldest infectious disease in the world. After pathogen recognition by innate immune cells a serie of intracellular events are triggered culminating at the translocation of NF-κB to the nucleus. The DNA binding region of NF-κB is crucial for the coding of inflammatory genes resulting in the production of many inflammatory mediators implicated in the host defense against mycobacteria infection. In vitro studies showed that the tiophenolic compound, thiophenacetamide (TAA) and its analogs have moderate to low activity against *M. tuberculosis* and low cytotoxicity against macrophages. The aim of this study is to evaluate the binding mode of TAA and analogues within NF-κB, to study its dynamical behavior and the consequences in vitro of its interaction. These results will help us propose ways to interfere with the host immune response, in order to eliminate the infection. To achieve this goal, we used molecular docking and molecular dynamics methodologies for the in silico assays. For our in vitro studies we used an experimental model of macrophages infection. The binding pocket prediction retrieved eighteen possible cavities in the protein. Docking simulation recognized a particular pose for TAA and their analogues using AutoDock Vina 1.1.2 and DockThor server. TAA interacted mainly with Tyr57 and Val142 residues, as its analogues did with Tyr57, which composes of the NF-κB active site. The stability of cavities along time was checked, and variations in volume were detected. The ligand absence affected the average lifetime of the cavities. Molecular dynamics simulations revealed that the presence of DNA in the protein helps stabilize the complex (RMSD = 0,5 Å). The simulation time was 50 nanoseconds using NAMD program. As our data in silico demonstrated a highly possibility of interaction between NF-κB and TAA, we analyzed whether TAA would have an immunomodulatory action on macrophages infected with *M. bovis*- BCG. It was observed in the supernatant of these cells a decreased in the release of TNF-α, IL-6. In addition a qualitative assessment of the nuclear translocation of NF-κB demonstrated that the TAA was able to inhibit it nuclear translocation. Molecular docking methodology suggested that TAA locates in cavities predicted by online servers. Also, molecular dynamics suggested that the DNA promotes stability to NF-κB while the absence of DNA promotes protein instability. In addition our in vitro results suggest a reduction on the inflammatory response caused by the infection.

Supported by: Fiocruz, Capes, FAPERJ.

Identification of druggable binding sites in ribose-5-phosphate isomerase of *Trypanosoma cruzi*

RF Soares¹, ACR Guimarães², ER Ca arena¹

¹ Computational Biophysics and Molecular Modelling Group, PROCC, Fiocruz

²Functional Genomic and Bioinformatics Laboratory, IOC, Fiocruz

Diseases caused by the trypanosome family members are a major public health problem in tropical and subtropical regions of developing countries, particularly in Brazil. The World Health Organization estimated that approximately 10 million people are infected with *Trypanosoma cruzi*, the etiologic agent of Chagas disease, with most cases found in Latin America. Unfortunately, there are no vaccines to control Chagas disease, and the two currently available drugs, nifurtimox, and benznidazole are inadequate for several reasons. For instance, these drugs present significant toxicity, act only in the acute phase of the infection, and some strains of the parasite have developed resistance to the available treatment. Hence, the search for new treatment strategies for Chagas disease is crucial to the development of more efficient drugs. The enzyme ribose 5-phosphate isomerase (R5PI) is an interesting molecular target. The R5PI enzyme is part of the pentose phosphate pathway, and its role is to protect the parasite against oxidative stress and production of nucleotides and NADPH precursors. Despite presenting functional resemblance with its human counterpart (HsR5PI), R5PI shows differences in its primary and tertiary structures. In this work, we applied a structure-based rational design approach, which included the search for cavities on the surface of R5PI and the analysis of the potential druggability of these cavities. Our purpose here is to investigate the possibility of agonizing or antagonize the protein through another binding site not yet described. A 100ns molecular dynamic simulation of R5PI (PDB ID 3K70) without the D-ribulose-5-Phosphate (substrate) showed that Glu 121 and His 23 undergo conformation changes transiently to prevent the formation of the catalytic pocket reducing its druggability. From the simulation, we clustered the structures in 3 groups and, afterward, chose the most representative one to search for potential allosteric sites. From the analysis, ten possible allosteric sites emerged, but only 3 presented high druggability according to the Pockdrug server, with values ranging from 0.72 to 0.84 and associated volumes close to 450 Å³. However, only two were stable concerning druggability and can be used in virtual screening studies to evaluate allosterically. Although the majority of inhibitor candidates act in the active site of an enzyme, in this work we searched for other potential binding sites where ligands could bind and act indirectly by provoking conformational changes in the protein to the point of altering some biological properties. The results obtained from our combined methodology may help in the development of alternative therapies against Chagas disease.

VERMONT: A tool for mutation visualization

Sócrates S. Araújo Jr.¹, Samuel S. Guimarães¹, Alexandre V. Fassio²,
Raquel C. de Melo-Minardi², Sabrina de A. Silveira¹

¹*Universidade Federal de Viçosa*, ²*Universidade Federal de Minas Gerais*

Mutations are events that occur naturally due the evolution, changing the sequence of residues, which can possibly affect protein structure and function. Thus, an important open problem in Bioinformatics is to understand how these specific mutations in protein residues can affect or not on protein function. To tackle this problem, VERMONT (Visualization Mutation Tool) was proposed in the contest of IEEE BioVis 2013, where it received the Biology Experts Pick award and its paper was published on BMC Proceedings journal. At that time, VERMONT was a static tool that allowed users to study the impacts of a mutation on a specific dataset provided in the contest. Now, we present a generic version of VERMONT, which allows users to set up their own set of proteins to be analyzed. To start the analysis user can provide his/her own files in .pdb format, or the user can choose to get files directly from the Protein Data Bank (PDB). Next, a structural alignment of all protein family is computed using MultiProt and then the similar sequences are grouped to help visualization of the residue conservation on the dataset with the Expectation Maximization algorithm. Then, we modeled the protein structures as graphs in two different levels of granularity to compute contacts through the Delaunay triangulation: (i) as atomic graphs, where each node is a protein atom and each edge represents the interactions between atoms and (ii) as residue graphs, where each node is a residue and each edge represents the interactions among a residue pair. In both levels, the nodes and edges are labeled with their physicochemical properties. To analyze the protein networks, we used some measures of complex networks (degree, closeness and betweenness) that helps us to determine how central and connected a node is in the network. Highly connected nodes can potentially cause more damage in the protein's function in case they suffer a mutation. We propose interactive visualization strategies to show all computed data, especially the contacts conservation all over the dataset, coupled with centrality measures and solvent accessibility for all residues, allowing users to get details on demand by clicking or passing the mouse over each residue.

This work has been supported by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) and Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG).

New annotation strategies: from sequence to 3D structure

Rafael Nicolay B. da Silva¹, Paulo José Miranda da Silva Iwakami Beltrão¹,
Manuela Leal da Silva¹

¹*Instituto Nacional de Metrologia, Qualidade e Tecnologia - INMETRO*

The procedures applied to biological data annotation, nowadays, presents a computational network developed for analyze all kinds of data. Commonly, annotation algorithms perform comparisons between the raw data from a sample against renowned biological databanks. Annotation strategies are divided into sequential and structural, both strategies can be applied to increase the reliability for an analyzed dataset. This study will be presenting the identification of new molecular targets from sequential annotation of the feces from *Bradypus variegatus* metagenome sample. The sequential annotation strategy consists in the use of a known-function sequence, derived from renowned databanks, which is related to the function we want to identify within the sample. We performed local alignment techniques, in order to obtain contigs with high ratio of specificity. Further, we extracted the function-related sequences with short lengths to identify annotated sequences which presents a high percentage of identity, suggesting the presence of similar sequences inside the sample. At last, we performed a reverse search procedure using the annotated sequence, from previous step, against the biological sample. In this procedure, we extracted new fragments, besides the one employed to search for the annotated data, and performed a reference-based assembly and annotation of a new protein. As results, we identified a potential new enzyme characterized as a Glicosil Hydrolase family 8. The PSIPRED software was applied to predict secondary structures, the BLASTp suite was employed to perform local alignment techniques against the Protein Data Bank. The results revealed two potential enzymes, characterized as Cellulose synthase from *E. coli*, for templates with PDBIDs 3QXQ and 3QXF, both with 94% of coverage and 87%/85% of identity, respectively. Further, we applied MODELLER to perform the comparative modelling for generate 200 candidate models. The 3D structures generated from both templates were validated through different parameters. The best model presented values for Ramachandran's plot most favored and disallowed residue regions as 96.3% and 0%, 0.157 Å for RMSD, -40830.668 for DOPEscore and 100% for GA341 score. The new strategy consists in the capability of automation for the whole annotation process, considering the reference-based assemble, the compilation of a new sequence and the creation of valid 3D models for further structural annotation, not described in the literature as an automated process yet. The next step consists in performing the structural annotation using ASAProt software (Automatized Structural Annotation of Proteins) and analyze the possibilities of experimental applications with the identified enzyme.

Financial support: CAPES and CNPq.

Metabolic changes of pathogenic and nonpathogenic Leishmania species during host cell infection by integration of mathematical models, quantitative proteomics and untargeted $^1\text{H-NMR}$

Mendes, T.A.O.^{1,2}; Souza, D.M.^{1,3}; Cardoso, M.S.¹; Machado, A.R.T.⁴; Ferreira, S.R.¹; Almeida, L. V.¹; Marques, A.F.¹; Pimenta, L.P.S.⁴; Filho, J.D.S.⁴; Nakayasu, E.S.⁵; Fujiwara, R.T.¹; Patil, K.R.²; Bartholomeu, D.C.¹

1- Departamento de Parasitologia, UFMG – Brazil, 2- Structural and Computational Biology Unit, EMBL – Germany, 3- Departamento de Patologia Clínica, COLTEC – Brazil, 4- Departamento de Química, UFMG – Brazil, 5- Bindley Bioscience Center, Pacific Northwest National Laboratory – USA

The current treatment of visceral and cutaneous leishmaniasis is based on a very limited number of drugs with variable efficiency and many side effects. Thus, the discovery of new drugs and parasite targets is considered a priority strategy to disease control. During infection of mammalian macrophages including humans, Leishmania species changes morphology and biochemical profile in a process named amastigogenesis and blocking key enzyme of this process might interrupt the infection. The main goal of this work is the identification of important enzymes to metabolic network changes during in vitro amastigogenesis by comparative analysis of two visceral Leishmania species (*L. donovani* and *L. infantum*), one cutaneous disease (*L. major*) and one non-pathogenic specie (*L. tarentolae*). The fist step was the reconstruction and simulation of metabolic model based on the proteins encode in the genome of four Leishmania species. Since the disease phenotype depends of differential genome content and differential expression, we performed time-course quantitative proteomic analysis for all species during axenic amastigogenesis. The proteomic data were integrated to mathematical metabolic models to predict metabolite concentration changes during the process. The congruence of model was evaluated by comparison of predict metabolite concentration changes with experimental values obtained by untargeted $^1\text{H-NMR}$ metabolomics with Pearson correlation coefficient between 0.66 and 0.79 (p-value < 0.0001). Interestingly, visceral parasites have the highest metabolic changes followed by cutaneous disease organism. Specifically, pathway associated to protection to oxidative macrophage response and nutritional requirement of intracellular parasite stage were enriched in infective species compared to apathogenic *L. tarentolae*. We identified the proteins homoserine kinase and trypanothione synthase as key enzymes responsible to control metabolite concentration in important pathways associated with *Leishmania* infection. The importance of these proteins have been validated using mutant parasites super-expressing each enzyme and we have been identified small molecules able to block their functions. Acknowledgements: CAPES, CNPq and FAPEMIG.

Prospecting novel proteins from *Deinococcus radiodurans*: a model for putative heat shock proteins

Ricardo Valle Ladewig Zappala^{1,2}, Pedro Geraldo Pascutti^{1,2},
Manuela Leal da Silva^{2,1} e Claudia A. S. Lage¹

¹Instituto de Biofísica Carlos Chagas Filho, Universidade Federal do Rio de Janeiro (RJ),

²Diretoria de Metrologia Aplicada às Ciências da Vida, Instituto Nacional de Metrologia,
Qualidade e Tecnologia (RJ)

The Deinococcaceae group comprises some of the robust known extremophilic bacteria. Attempts have specially focused on responses against extreme doses of gamma radiation, to explain survival mechanisms of *Deinococcus radiodurans* against simultaneous stresses, as desiccation and heat. *D. radiodurans* has many defensive mechanisms, and transcriptomes already made in response to gamma radiation and desiccation revealed that some genes were transcribed to proteins of undefined functions, while others have never been expressed under those conditions. Therefore, it is expected that such genes with obscure function can code for novel resistance proteins to these extremophilic conditions. The present study aims to identify and perform function prediction for hypothetical, unique proteins of *D. radiodurans*, without similarity to any other known protein. A group of proteins expressed in *D. radiodurans* after gamma radiation was retrieved, which hypothetical functions were predicted by the best scores after Psi-BLAST alignments and CD-search. Information about the proteins was gathered through alignments against Uniprot and PDB databases. Using molecular modeling tools as I-TASSER, SWISS MODEL and MODELLER, 3D models were successfully built for 20 out of 26 hypothetical proteins and they were initially evaluated by Ramachandran's Plot and RMSD. The best models from were then submitted to structural classification on SCOP and CATH servers. That way, we were able to speculate about the function of some candidates, and generate other models whenever structural and sequential annotations disagree. Among the 20 analyzed proteins, one of the most interesting was the DR0491 gene product, showing 25% identity and 41% similarity covering 90% of the sequence correspondent to the *Escherichia coli* heat shock protein Hsp31. This may represent an essential role on catalysis of damaged proteins, as well as proper folding assistance on other unstable proteins. Moreover, the catalytic essential residues of the *E. coli* protein were also found at the correct position in our *D. radiodurans* modeled protein. Other 19 proteins seem to bind to nucleic acids, acting on metabolic regulation under severe conditions. This particular resistance toolbox with novel and exclusive proteins was referred as the "Black Box Genome of *D. radiodurans*". Additionally, our results reveal promising candidates for future biotechnological approaches.

Funding support – CAPES, CNPq

Definition and comparative analysis of the kinomes of *Leishmania infantum* and *L. braziliensis*

Joyce Villa Verde Bastos Borba¹, Arthur Carvalho Silva¹, Pablo Ivan Pereira Ramos², Nicholas Furnham³, Carolina Horta Andrade¹

¹LabMol - Laboratory for Molecular Modeling and Drug Design, Faculty of Pharmacy, Federal University of Goiás, Goiânia, GO, Brazil, ²Instituto Gonçalo Moniz (IGM), Fundação Oswaldo Cruz (FIOCRUZ), Salvador, Bahia, Brazil, ³Department of Infection and Immunity, London School of Hygiene and Tropical Medicine, WC1E 7HT, London, United Kingdom.

The parasites of genus *Leishmania* are causative agents of leishmaniasis, an endemic disease in 98 countries grouped as neglected tropical disease by the World Health Organization. There are only few drugs available for the treatment and they face issues such as toxicity, lack of efficacy, route of administration and emergence of resistant strains. Therefore, it is urgent to search for new drug targets in *Leishmania*. Protein kinases (PKs) are potential drug targets given their essential role in many biological processes. We performed a proteome-wide analysis of PKs of the species *L. infantum* and *L. braziliensis* using a refined bioinformatics pipeline. First, we classified the PKs from both species proteomes using Kinannoter software. Then, we added orthologous kinases from previously established kinomes of close organisms using the softwares OrthoMcl and OrthoVenn. We also curated our classification by constructing hidden Markov models (HMM) profiles of kinase groups from close organisms and searched through both species proteomes. Next, we performed a multiple alignment of the kinase domains and constructed a phylogenetic tree using the softwares MAFFT, Muscle and MEGA7. Then, the functional annotation of the predicted kinases was performed using Interproscan, KEGG and Gene Ontology terms. In order to find kinases that could be druggable targets, we searched for their essentiality at TriTrypDB and selected proteins associated with lethal phenotype. As a result, a total of 211 and 204 PKs were identified in *L. infantum* and *L. braziliensis*, respectively. The eukaryotic (ePKs) were classified into six of the nine major kinase groups and many kinases could be classified into family and subfamily levels. The most representative groups were CMGC (n= 50/48) and STE (n = 42/41). The poorly representative groups were AGC (n = 13/11), CAMK (n = 23/22) and CK1 (n = 7/7). The comparison of the kinomes of *L. infantum* x *L. braziliensis*, *L. infantum* x *L. major* and *L. infantum* x *Homo sapiens* showed a range of sequence identity between 77-81%, 81-94% and 33-47%, respectively. When we searched for essential kinases with lethal phenotype, 3 kinases were found: a polo-like PK (LinJ.17.0770), which is involved in signal transduction, cellular growth and apoptosis; an aurora kinase (LinJ.28.0550), involved in mitosis; and a casein kinase (LinJ.35.1030), involved in signal transduction. In conclusion, this bioinformatics pipeline provided the definition of *L. infantum* and *L. braziliensis* kinomes and could be useful for further studies of protein kinases as drug targets for antileishmanial drug design.

Development of cruzain selective inhibitors by structure based virtual screening

Viviane Corrêa Santos, Rafaela Salgado Ferreira

Universidade Federal de Minas Gerais

Cruzain is the major *Trypanosoma cruzi* cysteine protease and is related to parasite nutrition and host cell invasion. Its inhibition is shown to decrease parasite infection in animal models but no medicine has been developed yet. Cruzain is homologous to human cathepsins, so, selectivity can be a challenge when developing drugs against this enzyme. In order to develop a selective inhibitor to cruzain we propose a virtual screening with docking. Glide (Schrodinger) was the chosen software, it has 3 precision levels that increases in accuracy in pose prediction, true positives recognition and computational cost (HTVS, SP and XP, in this order). The crystals structures we applied in our docking protocols have the following PDB IDs: 3KKU (cruzain), 3AI8 (human cathepsin B) and 1MHW (human cathepsin L). All the enzymes were prepared with Protein Preparation Wizard from Schrodinger. We downloaded the Leads Now compounds from ZINC and selected compounds that differ from each other with a Tanimoto cutoff of 0.9. A diversity set of 372,632 compounds was prepared with LigPrep (Schrodinger) and they were submitted to Glide HTVS against cruzain. Top 10% compounds were submitted to Glide SP against cruzain and top 10% were filtered based in the presence of some interactions between ligand and receptor, in this step we retrieved 2,025 compounds. These were submitted to Glide SP against both cathepsins and bottom 10% (489 molecules) were submitted to Glide XP against all the enzymes. Molecules were visually inspected and selected based in the occupancy S2 pocket of some portion of ligand; hydrogen bonding pattern; chemical diversity of compounds and purchasability of supposed hits. We selected 9 molecules to be purchase and test against cruzain and human cathepsins in order to check its activity against cruzain and its selectivity towards its humans homologous.

Supported by: CNPq, CAPES and FAPEMIG

Protein Folding by Generalized Simulated Annealing and Molecular Dynamics Methods

Tácio Vinício Amorim Fernandes and Pedro Geraldo Pascutti

*Instituto de Biofísica Carlos Chagas Filho, Universidade Federal do Rio de Janeiro,
(Rio de Janeiro, Brasil)*

Protein 3D structures are obtained basically by X-ray Crystallography and Nuclear Magnetic Resonance. However, due to experimental limitations and the high costs involved in those techniques, determining these structures is often a demanding challenge. This has led to an enormous and growing gap between the number of known sequences and determined structures. In this sense, theoretical and computational studies have made possible to increase the comprehension of the factors that lead to a polypeptide folding into its native 3D state. In general, it is assumed that the native structures are found in the global free energy minimum and the information to achieve it is stored in amino acid sequence. The objective of this work was to develop a simulation methodology based on Generalized Simulated Annealing (GSA) and Molecular Dynamics (MD) in implicit solvent, for protein structure prediction. In order to validate this study, we used a set of 65 protein models with lengths ranging from 10 to 60 residues. We applied the GSA in the search for a conformation into energy folding funnel, and then MD to refine the models. The results show that the proposed protocol is able to find models very close to the native structures determined experimentally. For about 57% of the sequences analyzed, we found models with less than 3.0 Å of deviation from the experimental structure, which is considered a high quality prediction. Furthermore, over 70% of the generated models showed deviations below 4.0 Å, and 87% less than 5.0 Å, which are considered good results in the literature. In general, our results showed that the optimization method with GSA, from extended conformation combined with further MD refinement, is a promising strategy to find native states.

Financial support FAPERJ, CAPES and CNPq

Mutation Analysis for *AgrC* from *Staphylococcus aureus*

Samuel da Silva Guimarães, Danielle Mendes Silva, Mônica Pacheco Silva, Pedro Marcus Vidigal, Andréa de Oliveira Barros Ribon, Sabrina de Azevedo Silveira

Universidade Federal de Viçosa

Staphylococcus aureus is one of the main pathogens of bovine mastitis. The analysis of SNPs on the genomes of four strains of *S. aureus* associated with mastitis showed the presence of 6 variations in 5 positions in the sequence of the *agrC* protein. Thus, it was hypothesized that these variations could be related with different manifestations of the disease. We searched for the reference sequence of *agrC* in Protein Data Bank (PDB) and found the entry 4BXI.A, which comprehends just part of the sequence, with 153 residues (278-430). A new search was made in the PDB to find structures similar to 4BXI.A, resulting in a set of 82 different entries, grouped by 40% of sequence similarity. A pairwise structural alignment was performed using the MultiProt to align each of the structures against 4BXI.A. A visual representation for this alignment was generated using the CINEMA color scheme, so that each sequence is represented by a line and each column corresponds to an alignment position. Also, we used the Expectation Maximization (EM) algorithm to group similar sequences, so that these similar sequences appear next to each other, which helps the user to detect trends and exceptions in the data. Next, the interactions were modeled as graphs in which nodes represent residues and edges represent interactions between residues. To calculate the interactions we used the Voronoi diagram followed by the Delaunay triangulation and we labeled nodes as positively charged, negatively charged, aromatic, hydrophobic, donor or acceptor. The edges were labeled according to a distance criteria and the type of edges as hydrogen bond, aromatic stacking, hydrophobic, repulsive and salt bridge. From the set of proteins modeled as graphs, some centrality measures that are commonly used in complex networks were calculated using the iGraph package from R, each of them providing a different perspective of centrality. This strategy enabled us to identify 2 positions on the *agrC* sequence where SNPs can potentially impact on protein structure and should be further studied to evaluate possible connections to bacterial virulence mechanisms.

Financial support: CAPES, CNPq, FAPEMIG.

nAPOLI: a web tool for protein-ligand interactions analysis

Alexandre Fassio^{1,2}, Sabrina Silveira³, Rafaela Ferreira², Raquel de Melo-Minardi¹

¹*Department of Computer Science, UFMG;* ²*Department of Biochemistry and*

Immunology, UFMG; ³*Department of Informatics, UFV*

Elucidating the mechanisms involved in the molecular recognition and which forces contribute to the recognition is a central problem in biology, since the interactions between two molecules are extremely important to biological systems as a whole and are very toilsome to be predicted even for small molecules. Nowadays, many interesting tools exist to analyze protein-ligand interactions, however, none of them presents large scale statistical, visual and interactive analysis. Thus, users must perform their comparison manually, that is completely infeasible. Therefore, we propose an easy and intuitive tool through visual strategies to depict the patterns and the types of interactions established between proteins and their ligands. In order to achieve it, we propose nAPOLI (Analysis of PrOtein Ligand Interactions), an interactive web tool to study the protein-ligand interactions in large scale by using visual strategies and statistical analysis. Since from the first version, many improvements were added to nAPOLI as well as new features. nAPOLI as a server allows users to submit their own dataset or to use PDB files from the PDB.org database. Moreover, many interactive functionalities were designed in this version like, for example, the 3D and 2D-view of protein-ligand complexes, comparison between clusters and download of data. Additionally, we performed improvements in the method used to calculate interactions which contributes to detect accurately such interactions. Some of them are: hydrogen bonds are now detected by using angle; atoms properties were better defined by using properties like pH and atoms charge; and the cutoff used to set interactions were also refined to improve nAPOLI results; etc. Finally, we compared nAPOLI to other two tools that is CREDO and LIGPLOT. Through this comparison we could show that nAPOLI presents accurate results regarding to these tools. Thus, nAPOLI showed to be very accurate and useful as it shed some light on the problem of discovering what is conserved in a set of ligands and which interactions they establish with a receptor.

Supported by: Brazilian agencies Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG), Financiadora de Estudos e Projetos (FINEP) and Pró-Reitoria de Pesquisa da Universidade Federal de Minas Gerais.

Characterization of an antimicrobial peptide from eggplant leaves as an inhibitor of carboxypeptidase

Maria Cristina Baracat-Pereira¹, Victor Dose Lage de Almeida¹, Hebréia Oliveira Almeida-Souza¹, Maura Vianna Prates², Marcelo Porto Bemquerer², Tiago Antônio Oliveira Mendes¹

Department of Biochemistry and Molecular Biology, Institute of Biotechnology Applied to Agriculture, Universidade Federal de Viçosa, Viçosa-MG, Brazil; Brazilian Agricultural Research Corporation (EMBRAPA), Brasília-DF, Brazil.

The constant use of antibiotics to control diseases resulted in the selection of resistant pathogens. Antimicrobial peptides (AMPs) represent a primitive defense mechanism of plants, still poorly understood, and represent sources of defense agents with new mechanisms of action and the advantage of being non-inducer of microbial resistance. Solanaceae such as peppers, eggplant, tomatoes and potatoes are promising as AMPs sources for biotechnological applications, which is the aim of our research group. An antimicrobial peptide was purified from leaves of eggplant (*Solanum melongena*) in the Laboratory of Proteomics and Protein Biochemistry at Federal University of Viçosa (Viçosa-MG). The peptide presented 4,140 Da and six cysteine residues linked by three disulfide bonds. The 30 first residues were sequenced by automatic Edman sequencing and showed no significant similarity to other peptides described in UniProtKB/Swiss-Prot, RefSeq and PDB (by BLASTp). In the Eggplant Database, which contains the complete genome of the eggplant, a sequence encoding 59 residues was identified showing identity with 28 of the 30 residues already obtained, which allowed completing the sequence of 37 residues of the eggplant peptide. ClustalOmega showed only two different residues in eggplant peptide when compared to the peptide of 59 residues (19Val-Ile and 21Gln-Trp). PSI-BLAST identified three peptides with high identity in the sequence, all obtained from Solanaceae: for Ref-Seq, a pepper (*Capsicum annuum*) peptide with 86 residues, and for NCBI and PDB, a tomato (*Solanum lycopersicum*) peptide (PDB-2HLG) and a potato (*Solanum tuberosum*) peptide (PDB-1h20), both with 39 residues. JPred and PSIPRED showed an alpha-helical region (residues 16-27) and a beta structure (residues 31-35), similarly to the tomato and potato peptides. The tomato and potato peptides also contain three disulfide bonds with cysteine residues in equivalent positions to the eggplant peptide. Pfam indicated the presence of a carboxypeptidase inhibitor (CPI) domain, which is also present in tomato and potato peptides. Thus, the peptide of *S. melongena* was named CPI-SMEL. Molecular modeling of CPI-SMEL (37 residues) by similarity using PHYRE² indicated these tomato and potato peptides as molds, and was able to position two of the three disulfide bonds, missing the bind near to the C-terminus. The CPI activity will be determined in vitro to confirm the CPI-SMEL function, following studies aiming the use in biotechnology. (FAPEMIG, CNPq, CAPES, FINEP, NuBioMol, BIOAGRO).

Proteomic analysis of seminal plasma from stallions (Mangalarga Marchador) influenced by seasonality

Renato Lima Senra, João Gabriel da Silva Neves, Marcos Jorge Magalhaes Júnior, José Domingos Guimarães, Maria Cristina Baracat-Pereira

Department of Biochemistry and Molecular Biology, Institute of Biotechnology Applied to Agriculture, Universidade Federal de Viçosa, Viçosa-MG, Brazil

The horses are seasonal polyestral animals, and the reproductive activity is primarily regulated by the photoperiod. The circannual differences affect also the composition and content of seminal plasma, including proteins in the plasma that interact with the surface of the spermatozoa and modify the characteristics of their membrane. Proteins in seminal plasma have been shown as involved in stages of the fertilization process as the establishment of sperm reserves in the oviduct, the modulation of the sperm capacitation, and the interaction between gametes. Proteomics of the seminal plasma from stallions using two-dimensional electrophoresis (2-DE), mass spectrometry and bioinformatics analysis may contribute to the understanding of biochemical factors involved in fertility and semen quality. The aim of this study was to evaluate the total proteome of semen from stallions in reproductive age, and detect changes in the protein profile between seasons, aiming to evaluate the physiological consequences for equine reproduction process. Semen samples were harvested, fractions were reserved for andrological analysis, and sperm were removed, recovering the seminal plasma by centrifugation. For the 2-DE first dimension, the seminal plasma samples were subjected to the isoelectric focusing using IPG strips (24-cm) with pH gradient from 3 to 10 (EttanIPGphor System 3, GE Healthcare, USA). For the second dimension, proteins were separated in SDS-PAGE gel 14.5% T and stained with coomassie blue. Gels were scanned and analyzed using Image Master 2D Platinum 7.0 software (GE Healthcare, USA). The spots of all proteins from gels were manually localized and excised, proteins were reduced, alkylated and trypsinized, following sample analysis in a MALDI-TOF/TOF (Ultraflex III - Bruker Daltonics) mass spectrometer (by MS and MS/MS). We found a set of proteins showing differential abundance in the different seasons. The samples were identified using the MASCOT DAEMON v.2.0 and Peaks Studio 7.5 softwares, and the UniProt, SwissProt, Equidae databases. The data are being analyzed with the aid of bioinformatics tools associated with ExPASy platform and String. Our results indicated the presence of a group of kallikreins in the seminal plasma in breeding season. It is expected the identification of metabolic pathways and biological events that will allow to suggest procedures for the improvement of equine reproduction process. (Support: FAPEMIG, CNPq, CAPES, FINEP, NuBioMol, BIOAGRO).

GA^SS-WEB: a web server for identifying enzyme active sites based on genetic algorithms

João P. A. Moraes¹, Douglas E. V. Pires², Raquel C. de Melo-Minardi^{3,4}, Gisele L. Pappa^{3,4},
Sandro Carvalho Izidoro¹.

¹Advanced Campus at Itabira, Universidade Federal de Itajubá, MG, ²Fiocruz Minas, MG,

³Computer Science and ⁴Department of Biochemistry and Immunology, Universidade Federal de Minas Gerais, MG.

Structure-guided methods have been proposed over the years to infer protein function based on active site similarity. Given an active site template, these methods use different mathematical modeling and searching procedures to match the template to a given set of proteins. Many of the current available methods present, however, limitations such as performing only exact matches on template residues (not accounting for conservative changes), pruning the search space using ad-hoc procedures, and not being able to find inter-domain active sites. In order to tackle these problems, we have proposed GA^SS (Genetic Active Site Search), a search method based on genetic algorithms that aims to cope with the aforementioned issues. Here we propose a user-friendly web server implementing the method's capabilities, called GA^SS-WEB. GA^SS-WEB can be used under two different scenarios: (a) given a protein of interest, to try to match a set of specific templates (*i.e.*, known active sites); or (b) given an active site template, looking for it in a database of protein structures. The method has shown to be very effective on a range of experiments. Based on the Catalytic Site Atlas (CSA) annotation, it was able to correctly identify >90% of the cataloged active sites. It also managed to achieve a MCC of 0.63 on the Critical Assessment of protein Structure Prediction (CASP 10) data set, ranking fourth among 18 methods. We believe GA^SS-WEB would be an invaluable tool to aid in active site search and protein function prediction of newly discovered proteins.

Supported by: FAPEMIG, CAPES and CNPq.

In silico structural studies of phospholipases A₂ inhibitors from snake blood

Carlos A. H. Fernandes¹, Fábio F. Mattioli¹, Liza Figueiredo Felicori², Consuelo L. Fortes-Dias², Marcos R. M. Fontes¹

¹Departamento de Física e Biofísica – Instituto de Biociências de Botucatu – SP, UNESP,

²Departamento de Bioquímica e Imunologia – UFMG, Brasil; ³Laboratório de Enzimologia Aplicada, Diretoria de Pesquisa e Desenvolvimento, Fundação Ezequiel Dias, Belo Horizonte – MG, Brasil.

Several snake species possess endogenous phospholipase A₂ inhibitors (PLIs) in their blood plasma, which their primary role is protection against an eventual presence of toxic phospholipase A₂ (PLA₂) from their venom glands in the circulation. These inhibitors have an oligomeric structure of, at least, three subunits and have been categorized into three classes (α , β and γ) based on characteristic structural features. In the present work, we constructed *in silico* models of the three classes of PLIs from South American snakes by threading modelling using Phyre2 server and molecular dynamics simulations using GROMACS v.4.5.3 software in GROMOS 96 53a6 force field; starting from sequenced or deduced amino acid sequences. The model of α PLI from *Bothrops alternatus* (named BaltMIP) presented the typical features of C-type lectin domains under monomeric configuration: an α -helical neck and carbohydrate recognition domain (CRD). We also constructed the *in silico* model of BaltMIP trimer by C_{α} atom alignments between the final α PLI model and the monomers of the homologue trimeric human lung surfactant protein D (SP-D) and simulated annealing simulations. Structural analysis of the BaltMIP trimer confirmed that α -helical neck is essential for trimer stabilization. Besides, CRD domains form a negatively charged central pore, which is actually the binding site for acid PLA₂. The β PLI model, based on translated aminoacid sequence of a β PLI transcript isolated from *Bothrops jaracussu* liver, presented the characteristic tandem leucine-rich repeats (LRRs) in its structure. These LRRs are rich on positively charged residues that could constitute the binding site for basic PLA₂. Finally, the γ PLI model from *Crotalus durissus terrificus* (named CNF, standing *Crotalus* neutralization factor) displayed well-defined three-finger domains in its tertiary structure. Besides, structural analysis of CNF *in silico* model combined to experimental data showed that tyrosine residues could play an important role in the oligomerization of CNF.

Funding support: FAPESP, FAPEMIG, INCTTox, CAPES and NCC/GridUNESP.

*In silico structural studies of Replication Proteins A1 and A2 from trypanosomatids (*Leishmania* and *Trypanosoma*)*

Carlos A. H. Fernandes¹, Fábio F. Mattioli¹, Raphael S. Pavani², Marcos R. M. Fontes¹, Maria C. Elias², Maria I. N. Cano³.

¹Departamento de Física e Biofísica - Instituto de Biociências de Botucatu – SP, UNESP, Brasil, ²Laboratório Especial de Toxinologia Aplicada – Instituto Butantan – SP, Brasil,

³Departamento de Genética - Instituto de Biociências de Botucatu – SP, UNESP, Brasil.

Replication Protein A (RPA), the major single stranded DNA binding protein in eukaryotes, is composed of three subunits (RPA-1, RPA-2 and RPA-3) and is a fundamental player in DNA metabolism, participating in replication, transcription, repair, and the DNA damage response. However, these proteins were not yet characterized yet in trypanosomatids, among which are causative agents of some neglected tropical diseases, such as leishmaniasis (caused by species of *Leishmania* genus) and Chagas disease (caused by *Trypanosoma cruzi*). At the present work, we constructed *in silico* models of RPA1 and RPA2 from *Leishmania amazonensis* (LaRPA-1 and LaRPA-2) and *Trypanosoma cruzi* (TcRPA-1, TcRPA-2) by threading modelling and molecular dynamics simulations. Both LaRPA-1 and TcRPA-1 lack the N-terminal 70N domain, that is present in RPA-1 from higher eukaryotes; but present three subsequently OB-fold domains (OBF-1, OBF-2 and OBF-3) which are homologous to the DNA binding domains A, B and C (DBD-A, DBD-B and DBD-C) of RPA-1 from *Homo sapiens* (HsRPA-1) and *Ustilago maydis* (UmRPA-1). However, comparative structural analysis between LaRPA-1 and TcRPA-1 *in silico* models and HsRPA-1 and UmRPA-1 crystal structures revealed that trypanosomatids RPA-1 present different binding modes of single stranded DNA (ssDNA). Whereas in HsRPA-1 and UmRPA-1 structures all DBDs are able to bind ssDNA, in TcRPA-1 model only OBF1 and OBF2 are able to interact with ssDNA. Interestingly, in LaRPA-1 *in silico* model, OBF1 is the unique protein region that interacts with the nucleic acid. Regarding RPA-2, LaRPA-2 and TcRPA-2 present an OB-fold domain and a C-terminal winged helix-loop-helix domain (wHLH) homologous to RPA-2 from higher eukaryotes. *In silico* models of OB-fold domain from LaRPA-2 and TcRPA-2 adopt a similar tertiary structure conformation compared to the HsRPA-2 and UmRPA-2 crystal structures. However, the trypanosomatids RPA-2 *in silico* models present a ten residue insertion rich of flexible residues located in the neighborhood of the DNA binding channel. In the molecular dynamics (MD) simulation, this region presented a high root mean square fluctuation of the main chain, adopting multiple positions during 50 ns of MD simulation, even blocking the DNA binding channel. These data suggest that the DNA binding site of trypanosomatid RPA-2 is more structurally unstable than their homologues in higher eukaryotes, causing changes in the DNA binding affinity trypanosomatids RPA-2.

Funding support: FAPESP and NCC/GridUNESP.

Mutational Analysis of Human K-ras G12C and Design and Molecular Docking of ARS-853 derivatives

João Marcos Galúcio¹, Cássio Figueira¹, Kauê Santana da Costa¹

Universidade Federal do Oeste do Pará

K-Ras is a Ras family protein primarily involved in the regulation of cellular proliferation, cell differentiation and apoptosis, alternating between an inactive form bound to GDP and an active GTP-bound form. K-ras mutations occur in approximately 30% of all cases of human cancer. However, despite of intensive research, relevant therapies for cancers with mutations in this protein have not been developed yet. K-ras G12C is the most common variant in lung cancer and specific therapies for this oncoprotein are in initial stages of development. ARS-853 is a potent inhibitor that binds to the inactive form of K-ras G12C, preventing its activation. This work purposes to design molecules with optimized affinity and complementarity in relation to receptor and relate the inhibitor interactions to structural stability of K-ras G12C. The native and mutant K-ras structures were analyzed by alanine scanning, stability calculations and interaction energy estimation between GDP and proteins in the Fold-X program. Based on redocking, physical-chemical profile analysis and prediction of pharmacokinetic and toxicological properties, ARS-853 derivatives were designed and docked to Switch II cavity. Our results provided evidences that G12C mutation has a neutral effect on the stability of K-ras. Likewise, the alanine scanning showed that the residues G12 and C12 have neutral contribution in protein stability, but the presence of C12 induces a charge inversion in the electrostatic potential map of the Switch II cavity that can be critical to selective inhibition of mutant K-ras. The presence of ARS-853 generated a high increase on the total energy of K-ras G12C and the interaction between GDP and the oncoprotein with inhibitor was destabilized. Two molecules derived from ARS-853 (Moldock score -202.09) were obtained (Moldock scores -227.66 and -215.77, respectively) with high complementarity to receptor and both are chemically simplified compared to the reference inhibitor. ARS-853 and its derivatives conserved some non-covalent interactions: hydrogen bonds with E63, R68 and D69 and hydrophobic pi-alkyl interactions with C12 and V103. Overall, these results reinforce experimental evidences that ARS-853 inhibits K-ras G12C with high selectivity, altering the stability of the mutant form. We conclude that it is prototype with great potential for development of new anticancer drug against this protein.

Modeling and Molecular Docking of the largest subunit of the Ribulose-1,5-Bisphosphate Carboxylase/Oxygenase (RuBisCO) from Alkalinema sp. CACIAM 70d

James Siqueira Pereira¹, Andrei Santos Siqueira², Leonardo Teixeira Dall'Agnol², Juliana Simão Nina de Azevedo¹ e Evonnildo Costa Gonçalves^{2,4}

¹Laboratório de Biodiversidade Molecular – UFRA, Capanema, PA ²Laboratório de Tecnologia Biomolecular – UFPA, Belém, PA, ³Universidade Federal do Maranhão – UFMA, Bacabal, MA, ⁴Centro de Inovações Tecnológicas – IEC, Belém, PA

Ribulose -1,5- bisphosphate carboxylase/oxygenase (EC 4.1.1.39, RuBisCO) is the most abundant protein in the world and is considered the major enzyme involved in the photosynthesis process and can be found in most autotrophic organisms like photosynthetic bacteria, cyanobacteria, algae and plants. RuBisCO is classified in four distinct forms: Forms I, II, III and IV. The form I of RuBisCO is a hexadecameric protein structure with eight copies of both large and small polypeptides in an $(L_2)_4(S_4)_2$ structure codified by *rbcL* e *rbcS* genes, respectively. This form is the predominant RuBisCO found in nature and it is present in Cyanobacteria, algae and plants. To unravel the structure and function of this enzyme in Cyanobacteria, this study aimed to construct a three-dimensional model (3D) of the large subunit of a Cyanobacterium from the 'Coleção Amazônica de Cianobactérias e Microalgas' – LTB/UFPA. The amino acid sequence was obtained from a genomic study of cyanobacterium Alkalinema sp. CACIAM 70d isolated from superficial water of Tucuruí Hydropower Plant Reservoir, Pará State, Brazil. The mold selection was chosen using Blast tool included in the PDB database. The best identity with RuBisCO from the *Synechococcus* PCC6301 (PDB ID: 1RBL.A). The three dimensional structure was generated through Modeller 9.10 and subsequently validated by the Ramachandran plot, Verify3D, Anolea and the Root Mean Square Deviation (RMSD). Finally, Molegro Virtual Docking was used for an analysis of molecular docking (MD) to evaluate the substrate in the catalytic site fitting. The obtained structure showed 15 β -sheets and 19 α -helix. The Ramachandran plot showed 98.28% of residues within energetically favorable regions and 89.29% of residues showed positive value in the 3D-1D evaluation. Individual residues analysis done by Anolea resulted in a few regions with high energy and the obtained RMSD value was 0.194. The map of electrostatic potential revealed similarity between the molecules regarding to their charge distributions, with low electron density even to the active site region. The best conformation obtained in MD process showed MolDock and Rerank scores -124,222 and -91.5559, respectively, significantly similar to those obtained for the template that showed values -135.27 and -96.8823. Furthermore showed the main interactions already described, highlighting those with Lys167, Lys169 and His290 residues, as well as with magnesium ion. The highest structural conservation, including electrostatic charges and interactions presented by the obtained model, classifies it positively, been contributing to studies that aims to optimize the carboxylase activity of RuBisCO and cyanobacteria biomass exploitation.

GReMLIN: A graph mining strategy to infer protein ligand interaction patterns

Charles A. Santana¹, Fabio R. Cerqueira¹, Carlos H. da Silveira², Alexandre V. Fassio³, Raquel C. de Melo-Minardi³, Sabrina de A. Silveira¹

Universidade Federal de Viçosa¹, Universidade Federal de Itajubá² and Universidade Federal de Minas Gerais³

Interaction between proteins and ligands are relevant in many biological process. Such interactions have gained more attention as the comprehension of protein-ligand molecular recognition is an important step to ligand prediction, target identification and drug design. This work proposes GreMLIN, a strategy to search patterns in protein-ligand interactions based on frequent subgraph mining. Here, we investigated if it is possible to find patterns that characterize protein-ligand interactions in a set of selected proteins. These patterns can be key factors to understand and support the recognition molecular process. Moreover, if such patterns exist, we believe that they can represent an important step in the prediction of the protein-ligand interaction.

Our strategy models protein-ligand interfaces as bipartite graphs where nodes represent protein or ligand atoms and edges represent interactions among them. The nodes and edges are labeled with physicochemical properties of atoms and a distance criteria. A clustering analysis is performed on graphs to characterize them according their similarities and differences, and a subgraph mining algorithm is applied to search for relevant patterns on protein-ligand interfaces in each cluster.

We collected structural data of protein-ligand complexes in Protein Data Bank (PDB) to validate our strategy and show their applicability. There are two datasets: (i) the CDK (Cyclin dependent kinases) dataset that have 73 PDB entries with identical sequences coupled with different ligands; and (ii) the Ricin dataset with 29 PDB entries, which share sequence identity greater than or equal to 50% with ricin template 2AAI chain A. Both datasets have biological relevance, but with different characteristics. Our strategy was able to find frequent substructures with considerable cardinality in the protein-ligand interfaces in the CDK and Ricin datasets. We provide the results of our strategy for the test datasets in a prototype interactive tool to visualize and explore the patterns found in protein-ligand interactions. Also, we provide a schematic 2D graph representation of such interactions and a 3D representation of these interactions in a molecule viewer. Availability: <http://homepages.dcc.ufmg.br/~alexandrefassio/gremlin/>. Financial support: CAPES, CNPq, FAPEMIG.

Modeling MS native-state amide hydrogen exchange through structural and dynamical properties

Machado, LA, Rodrigues, ABM, Carvalho, LMF, Costa, MGS, Bastos, LS, Batista, PR

Programa de Computação Científica, Fundação Oswaldo Cruz, Rio de Janeiro.

Proteins play an important role in all biological processes. Nevertheless, to perform such functions, they depend on their structural and dynamical properties. Nowadays, to probe these properties, hydrogen/deuterium exchange (HX)-based methods are often used. Hydrogen atoms from protein surface are in continual exchange with solvent. Thus, using deuterated water, it is possible to probe selectively the deuterium incorporation for each residue/peptide, using mass spectrometry (MS). Computational methods such as normal mode analysis (NMA) are well suited to study protein dynamics since it describes protein collective motions. Last decade, several theoretical models based on protein structure were developed in order to explain HX reaction, but all fail when systematically tested. This present work aims to develop a statistical model using protein structural (e.g. number of contacts and hydrogen bonds) and dynamical properties to explain protein native state MS-HX data. We built two different linear models: *i.* structural and *ii.* structural+dynamical model, where we use the atomic fluctuations from NMA to represent protein flexibility. We next evaluate the root mean square error (RMSE), Akaike Information Criteria (AIC) and the Pearson correlation coefficient between experimental data and the fitted values. The model using only structural features was not able to efficiently explain the HX data. However, the inclusion of a dynamical variable enhanced the correlation between its fitted values and HX data. In conclusion, we showed the use of fluctuations from NMA in conjunction with structural variables from a single structure allows one to obtain the closest correlation with experimental data from MS-HX.

ACKNOLEDGEMENT: We thank CAPES, CNPq and FAPERJ for the financial support.

Comparison Between a Graph-Based Metodology and the LSQKAB to Check Proteins Similarities

Monteiro, O.M., Dias, S.R., Rodrigues, T.S.

Centro Federal de Educação Tecnológica de Minas Gerais

Proteins are macromolecules present in all living beings and perform important functions, such as maintenance of organs and tissues, cell differentiation, transportation, and other several tasks. Many proteins have their three-dimensional structure solved and stored in biological databases, such as Protein Data Bank (PDB). There are softwares working with informations extracted from PDBs, some of these tools, has the function of checking similarities between protein structures. An example is LSQKAB, which belongs to CCP4 package and has its operation based on Kabsch algorithm, a technique frequently used in bioinformatics. Among the outputs of this software, is possible to get a list of the distances between each atoms pairs compared. However, this algorithm performs various computational calculations and it's use is inefficient to compare one record against a dataset. The main goal of this work is develop a methodology based on graphs to represent a protein structure in order to verify the similarity more efficient than LSQKAB, maintaining the accuracy. The first experiments were performed with 16,383 disulfide bounds files extracted from PDB. Each record contains 12 atoms and their distance values are represented by coordinates "x", "y" and "z". In the proposed methodology, each file is represented by a vector of 144 positions. Each position is refers to the calculation of Euclidean distance between the atomic distances of the 12 atoms. Using the representation, the clustering K-means technique was applied. The accuracy of each group generated was verified by LSQKAB trough of comparisons between all records of the same group. It was found that the more compact clusters correspond a more accurate result from LSQKAB, on the other hand, more spread clusters correspond a less accurate result from LSQKAB. This indicates that the behavior of the technique is similar to LSQKAB, moreover, more efficient on time consuming. The project is supported by Capes, CNPq and Fapemig.

Metabolic pathway prediction of enzymes: a machine learning approach

Rodrigo de Oliveira Almeida¹, Guilherme Targino Valente²

¹Instituto Federal de Educação, Ciência e Tecnologia do Sudeste de Minas – Campus Muriaé – MG, ²Universidade Estadual Paulista “Júlio de Mesquita Filho” - Faculdade de Ciências Agronômicas – Botucatu - SP

For long time, enzymes properties have been applied on several areas, since pharmacy to food industry. Nowadays, to establish an appropriate enzyme function is hard due to lots of biochemical procedures to be required. Bioinformatics is also focus to define protein functions by homologies or structural analysis; however those strategies are not applicable for most of proteins. Since the advance of sequencers along the last years, biological data generated are increasing fast and nowadays it is necessary more efficient tools to analyze this high amount of data. Thereby, machine learn is an interesting tool to help analyze those big data. The present study aims to construct models able to predict the metabolic pathway of enzymes based only on amino acid sequence properties. Protein sequences from four fungi (*Agaricus bisporus*, *Aureobasidium subglaciale*, *Saccharomyces cerevisiae* and *Talaromyces stipitatus*) were downloaded from Uniprot (<http://www.uniprot.org/>) and high similarity sequences (99%) were removed using the software CD-Hit. Data mining from protein annotations were performed to split in enzymes or non-enzyme proteins, which are the input dataset. For each metabolism (aminoacids, co-factors and vitamins, drug response, glycan, lipid and nucleotides) it was constructed a positive and a negative dataset. Around 1,200 protein attributes were generated using the R packages “Peptides” and “protr”. Relevant attributes were selected using Weka software tools. After this process, all datasets were normalized, the positive dataset was undersampled and balanced datasets were done; after that each dataset was submitted to supervised training using Weka software to generate prediction models. It was used 6 classifier algorithms (J48, Random Forest, RepTree, Sequential Minimal Optimization, Voted Perceptron and Multilayer Perceptron) to generate models for each metabolism and final models were generated using the MetaVote or MetaVoteBagging. All scripts were written in R language and ran in parallel using a Shell script to improve the time of performance. The averages of correctly classified instances for training were 87.07, 90.63, 97.14, 94.71, 95.02 and 86.73% (metabolism of aminoacids, co-factors and vitamins, drug response, glycan, lipid and nucleotides, respectively). The final models were applied on 2,607 enzymes sequences with unknown metabolic pathway (from the same organisms) to classify them. Those models were able to assign metabolic pathways for most of unlabeled enzymes, which the results of prediction ≥ 0.7 of probability show a mean of 6.82, 9.90, 1.66, 18.15, 17.36 and 2.01% of enzymes classified in metabolism of aminoacids, co-factors and vitamins, drug response, glycan, lipid and nucleotides, respectively.

Funding Support: BIOEN FAPESP

*In silico Identification of common putative vaccine candidates against *Treponema pallidum*: A reverse vaccinology based approach*

Arun kumar Jaiswal^{a,b}, Sandeep Tiwari^a, Syed Babar Jamal^a, Vasco Azevedo^a, Siomar C. Soares^{b*}

^a*Institute of Biologic Sciences, Federal University of Minas Gerais, Belo Horizonte, MG, Brazil.*^b*Department of Immunology, Microbiology and Parasitology, Institute of Biological Sciences and Natural Sciences, Federal University of Triângulo Mineiro (UFTM), Uberaba, MG, Brazil.*

Sexually transmitted infections (STIs) are caused by a wide variety of bacteria, viruses, and parasites that are communicated from one human being to another primarily by vaginal, anal, or oral sexual contact. Syphilis is also a serious disease of sexually transmitted infection. Syphilis is caused by the bacterium *Treponema pallidum* subspecies *pallidum*. *Treponema pallidum* is a motile, gram-negative spirochaete bacterium, it can be transmitted both sexually and from mother to child, and it can invade virtually any organ or structure in the human body. The current worldwide prevalence of syphilis emphasizes the need for continued preventive measures and strategies. Unfortunately, effective measures are limited. In this study, we mainly focus on identification of vaccine targets and putative drug against Syphilis disease using reverse vaccinology and subtractive genomics. We compared 13 strains of *Treponema pallidum* keeping *Treponema pallidum* Nichols as reference genome. Furthermore, the orthoMCL software was used to predict the cluster of orthologous genes. CDSs shared by all species were considered to be part of the core genome. Considering human as a host, a set of 565 conserved non-hosts homologous proteins were identified. These conserved non-host homologous proteins were analysed using reverse vaccinology for antigenic properties of candidate vaccine, subtractive proteomics and modelomics approaches for drug target identification. Based on this analysis, we have classified 207 gene products as secreted proteins, putative surface-exposed proteins or membrane protein. A set of 26 cytoplasmic proteins constituting distinct quality model were selected as drug target for the bacteria. These proteins were considered as essential and non-host homologs, and have been subjected to virtual screening using two different compound libraries (extracted from the ZINC database and plant-derived natural compounds). The proposed drug molecules show favourable interactions, lowered energy values and high complementarity with the predicted targets.

T lymphocytes epitopes prediction to access immunological response from Chagas diseases patient samples

Toledo, C.B.B^{1,2}, Castro, J.T^{1,2}, Gazzineli, RT^{1,2}, Junqueira, C¹

¹*Centro de Pesquisa René Rachou – CPqRR/FIOCRUZ Minas;* ²*Universidade Federal de Minas Gerais – UFMG, Belo Horizonte – MG*

Chagas Disease, a pathology caused by the intracellular protozoan *Trypanosoma cruzi* tackles approximately 16-20 million people in Latin America and is responsible for about 13 thousand deaths per year. There are still few prophylactic measures and pharmacological treatments available. Vaccine development is being broadly assessed by the use of different antigen and delivery vectors. To succeed, it is important that the vaccine induce a strong immune response through the recognition of epitopes presented by the human leukocyte antigen (HLA) to the T lymphocytes, that latter will lead to an immunological memory. Trans-sialidase (TS) and Amastigote Surface Protein – 2 (ASP-2) are two proteins expressed by *T. cruzi* with high antigenic potential. Both antigens were extensively evaluated in murine models, however a more robust analysis for human immunogenic potential should be performed. To accomplish that, a protein epitope prediction was made considering the great variability existent in the population's HLAs. Three different bioinformatics programs were used: SYFPEITHI, Bimas and The Immune Epitope Database and Analysis Resource (IEDB), once each one worked with a different algorithm. The obtained results were compared, and further analysis was made of every available HLA present in at least two programs. The top five ranked epitopes for each HLA were selected, and the most frequent epitopes were identified in the proteins' amino acid sequence. Considering the great importance of the recognition and affinity of the immune system for the antigen used in vaccination, we intend to screen the selected peptides for the activation of T CD8+ response, measured by the production of IFN- γ upon peptide stimulation. Aiming at screening the peptides, PBMC from Chagas patients will be sensitized with *T. cruzi* total antigen then re-stimulated with the selected peptides. The data will enlighten the immunological profile of different HLAs under specific binding of the said peptides and open the perspective to pinpoint an immunogenic region of interest of both TS and ASP-2 in order to improve the immunization protocols for humans.

Supported by: FAPEMIG, INCTV and CNPq

Standardization of the ribosomal protein genes nomenclature in *Leishmania major*

Thaís Couto Laureano¹, Felipe Freitas de Castro², Angela Kaysel Cruz², Patrícia de Cássia Ruy²

¹*Biological Sciences Institute – Universidade Federal de Minas Gerais;* ²*Department of Cell and Molecular Biology – Ribeirão Preto Medical School, Universidade de São Paulo*

Ribosomal proteins are responsible for the composition of ribosomes and also may have extra-ribosomal functions such as involvement in replication, transcription, splicing and cellular senescence. In the protozoan parasite *Leishmania*, transcription is constitutive and ribosomal protein genes are present in two or more copies in the genome; it is admissible that the higher number of copies is important to increase the level of transcripts generated. The lack of a rational nomenclature system for ribosomal proteins causes confusion since identical names given to some of these proteins (from different organisms) are unrelated in structure and function. Ban and co-workers proposed a new system for naming ribosomal proteins that take into account the three domains of life (archaea, bacteria, and eukaryote). We were based in Ban's nomenclature system to standardize the *Leishmania major* ribosomal protein naming. The two initial letters in the name refer to the domains; the prefix "ae" for a ribosomal protein present in archaea and eukaryote and prefix "be" for ribosomal protein in bacteria and eukaryote. We extracted from the TriTrypDB (tritrypdb.org) ribosomal protein sequences of *L. major* and all annotated CDS (Coding DNA Sequence) searching for "ribosomal protein". A total of 176 genes annotated as coding for ribosomal proteins in *L. major* were obtained. These sequences were compared with ribosomal protein genes of five other species (*Escherichia coli*, *Halobacterium salinarum*, *Saccharomyces cerevisiae*, *Arabidopsis thaliana* and *Homo sapiens*), obtained in UniProtKB (Universal Protein Resource Knowledgebase). Blastx tool was used in the sequence similarity search in two different stages: 1-) identification of putative ribosomal protein genes unannotated in the *L. major* genome; 2-) comparison of ribosomal protein sequences between *L. major* and the above the mentioned five organisms. Our result indicates the presence of 55 ribosomal proteins in all domains, 50 exclusively from Eukarya, 60 in Archaea/Eukarya and 9 in Bacteria/Eukarya. A Perl script was developed to filter the blast results and define unannotated ribosomal protein candidates. After standardization of the nomenclature, we observed that the majority of ribosomal protein genes are present in multiple copies. Approximately, 76% of the genes are present as duplicates, either in tandem or different chromosomes; 12.7% are single copy genes; 8.3% of them are triplicates, and 2.8% are quintuplices. This standardization of the ribosomal proteins nomenclature in *L. major* is relevant for a well-established and clear group definition and allows accurate comparison of ribosomal proteins from a broad diversity of organisms.

Financial support: FAPESP and CAPES

In silico prediction and comparative studies on plant GPCRs

Natália F. Martins, Emmanuel Bresso, Priscila Grynberg, Roberto C. Togawa, Deisy X. Amora and Bernard Maigret

EMBRAPA Recursos Genéticos e Biotecnologia, Parque Estação Biológica, Brasília, DF, Brazil; CNRS, LORIA, UMR 7503, Lorraine University, Vandoeuvre-lès-Nancy, France

G-protein coupled receptors (GPCRs) is the largest family of membrane proteins, as key components of signal transduction pathways. They are activated by diverse ligands, including odorants, fatty acids, peptides and neurotransmitters, across cell membranes. The identification of novel G protein-coupled receptors (GPCRs) from genome analysis allowed the prediction of new important receptors in several species. Despite the current knowledge, GPCRs in plants are not well characterized if compared to animals. The availability of plant genomes allowed the prediction and characterization of diverse GPCRs as well as a comparative analysis of the GPCR structure-function relationship. In plants, the repertoire of GPCRs remains unclear. Therefore, in silico analyses of candidates from *Arabidopsis*, *Capsicum annuum*, *Glycine max*, *Manihot esculenta*, *Medicago trunculata*, *Oryza sativa*, *Phaseolus vulgaris*, *Populus tricocarpa*, *Ricinus communis*, *Solanum lycopersicum*, *Solanum tuberosum*, *Sorghum bicolor*, *Theodroma cacau*, *Triticum aestivum*, *Vitis vinifera* and *Zea mays* describes the comparative survey for GPCRs with a total amount of 997,435 proteins. Following a logic pipeline the proteins were filtered by length resulting in 487,012 proteins between 250 and 1000 AA. Another trimming round of candidates with seven transmembrane segments revealed 4,878 possible proteins which were, therefore, submitted to the GPCRpipe program. This software predicted 60 GPCRs candidates, ranging from none (*Capsicum annuum*) to 12 candidates (*Populus trichocarpa*) (median = 3 ± 2.90). Comparative studies based on structure, sequence, ontology and phylogeny showed a low diversity in genes if compared to humans GPCR's classification. The Neighbor joining dendrogram of the 60 predicted GPCR candidates grouped the proteins in three groups. The first are similar to A Rhodopsin like, the second has similarity with GPCRs with Lung seven transmembrane receptor family protein and the third one was identified as GCR1 orthologs. These findings may indicate that GPCR in plants can be involved in drought stress responses, specially related to cytosolic calcium concentration changes in response to several stimuli, including light, pressure, gravity, and hormones.

Bioinformatics approaches to identify, classify and prioritize protein kinases as drug targets in *Schistosoma japonicum*

Arthur de Carvalho e Silva¹, Joyce Villa Verde Bastos Borba¹, Pablo Ivan Pereira Ramos², Nicholas Furnham³, Carolina Horta Andrade¹

¹Laboratory for Molecular Modeling and Drug Design, Faculty of Pharmacy, Federal University of Goiás, ²Gonçalo Moniz Institute, Oswaldo Cruz Foundation (FIOCRUZ),
³London School of Hygiene and Tropical Medicine.

Schistosoma japonicum is one of the parasitic atworms that causes schistosomiasis, a neglected tropical disease responsible for 200,000 deaths annually and affects more than 600 million people in Africa, Asia and South America. In this context, the main goal of this work was to determine the kinase complement within the *S. japonicum* proteome through bioinformatics tools, while also identifying essential kinases for the parasite that could serve as drug targets candidates. The proteome of *S. japonicum* was downloaded from GeneDB database. Next, the Kinannote tool was employed to generate a draft kinase, identifying 165 protein kinases candidates within the *S. japonicum* proteome. In a third step, kinases that Kinannote was not able to identify and classify were searched by means of orthology mapping with respect to *S. haematobium* and *S. mansoni*. The kinomes of both species were extensively used to analyze the orthology groups generated via OrthoMCL and OrthoVenn software, an approach that allowed the grouping of proteins from the three organisms. Using this strategy, the number of identified protein kinases in *S. japonicum* increased to 221. Functional annotation was made using KEGG and Gene Ontology terms, allowing the inference of subcellular localization and biological processes that each protein could be involved. MAFFT and MEGA7 software were used to build a phylogenetic kinase tree to help the visualization of the relationships between the kinase groups of *S. japonicum*. As a result of the classification, 7 kinases of *S. japonicum* were exclusively assigned to groups level, 103 kinases were assigned to groups and families; and 111 could be further classified into subfamilies. Kinases of *S. japonicum* were classified into nine major kinase groups: AGC (n=27), CAMK (n=35), CK1 (n=8), CMGC (n=38), Other (n=36), RGC (n=3), STE (n=22), TK (n=32), TKL (n=12) and atypical (n=8). STRING webserver was then utilized to generate a protein-protein interaction network based on *S. mansoni* data to study essentiality of some *S. japonicum* kinases. OrthoVenn analysis revealed 166 clusters of shared protein kinases among *S. japonicum*, *S. mansoni* and *S. haematobium* with a range of 75-85% of sequence identity and 116 clusters containing human orthologues shared by the three species. In conclusion, the present pipeline allowed the elucidation of the *S. japonicum* kinase and provides insights for next steps in the context of anti-schistosomal drug discovery.

Data Mining for characterization of nanotoxicity in mitochondrial ion channels induced by carbon nanotubes

Luisa Cornetet¹, Michael Gonzalez-Durruthy², Adriano Werhli¹, Jose Maria Monserrat², Karina S. Machado¹

¹*Centro de Ciências Computacionais (C3), Programa de Pós Graduação em Computação (PPGCOMP), Universidade Federal do Rio Grande (FURG),*

²*Instituto de Ciências Biológicas (ICB), Programa de Pós Graduação em Ciências Fisiológicas, Universidade Federal do Rio Grande (FURG)*

Single Walled Carbon Nanotubes (SWCNT) have been largely studied by the scientific community due to their great potential in several areas, ranging from industrial applications to medical uses. On nanomedicine, these SWCNT can be used to inhibit or stimulate the cells processes responsible for the production of energy required for their survival. The organelle accountable for delivering energy to cells is known as mitochondria. Ion channels found on those organelles play an important role on this process. Since the malfunctioning of these ion channels may induce some toxicity for the cells, the study of the interaction between carbon nanotubes and the ion channels is very helpful on understanding how these nanomaterials may induce toxicity on organisms. Aiming at understanding how carbon nanotubes may induce toxicity on organisms, this work proposes to perform molecular docking experiments with different carbon nanotubes and some mitochondrial ion channels proteins. After the execution of these simulations, we are going to apply a knowledge discovery in databases (KDD) process to relate carbon nanotubes characteristics with the docking results. With the results of KDD we expect to characterize the nanotoxicity that carbon nanotubes may induce on mitochondrial ion channels. To perform the molecular docking experiments we are going to use Autodock Vina and a Framework for Virtual Screening. As the target receptors we consider distinct mitochondrial ion channel proteins, found on Protein Data Bank (PDB). Some ion channel proteins do not have structure deposited on PDB and for those we are going to model the structure by homology using the sequences found on MitoProteome. As ligands, we are going to use carbon nanotubes with different geometries (arm-chair, chiral and zigzag), and also different functionalization, with both carboxyl and hydroxyl groups, totalizing 134 carbon nanotubes. After executing the molecular docking experiments, we will preprocess all these generated data. Then we are going to apply data mining algorithms and analyze the results aiming at characterizing the toxicity in mitochondrial ion channels induced by carbon nanotubes.

Prediction of protein stability changes upon single point mutation using Ensemble Learning

Alex D. Camargo, Adriano V. Werhli, Karina S. Machado

Centro de Ciências Computacionais, Universidade Federal do Rio Grande - FURG

The analysis of the destabilizing, neutral or stabilizing impact of single point mutations in proteins may be extremely valuable to further refine the relationship between sequence, structure and function of proteins. For this reason, computational tools were developed to predict the impact of point mutations. These models imply in the use of different assumptions about the probabilities of amino acid substitutions. Moreover, these assumptions seek for an approximation of reality supported by better accuracy. The combination of computational and statistical methods with experimental techniques, e.g. deep sequencing and high precision stability measurements, provides many supporting approaches for the protein stability engineering albeit the inherent computational problems. Most computational methods calculate the $\Delta\Delta G$ (free energy difference) between a wild type protein and its mutant which is considered an indicator of the mutation effects. Therefore, when there are different competing approaches to this problem, an effort to determine the most accurate is inevitable. The best approach depends on the available data and prior knowledge of the expert. Thus, the adoption of approaches to produce a final result better than individual results was sought using Ensemble Learning, taking into consideration that the values resulting from its classification can add greater generality by consensus. In doing so this work aims at developing an ensemble method to combine the results of different methods for predicting the impact of point mutation in proteins. At this moment we are considering the tools: I-Mutant, CUPSAT, SDM, mCSM, DUET, iRDP and MAESTRO. The initial proposal uses the plurality vote, popular in such learning as the key factor in the set. The dataset used in the case study came from the selection of experimental data from the biological databases Protherm and Protein Data Bank (PDB) totaling 1775 mutations. In general, the predictions had a good accuracy compared to the experimental values. For example, the tools I-Mutant and CUPSAT obtained accuracy of 75.21% and 65.57% of the predictions, respectively. The result the proposed ensemble (plurality vote) was similar to the best individual method, reaching 73.07% accuracy. As future work we are going to apply different mechanisms of ensemble and we are going to develop a tool based on the proposed method.

In Silico Binding Site Analysis of E6 Oncoproteins from High-Risk European HPV variants.

E R. Tamarozzi¹, G. Monteiro¹, S. Giulietti¹

¹ University of São Paulo - USP, Ribeirão Preto, SP - Brazil

Cervical cancer is the most studied oncopathology associated with Human Papillomavirus (HPV), which is present in all the studied cases of the disease. HPV of type 16 is the most prevalent, representing 70% of cases of cervical cancer worldwide. The SNPs throughout the virus gene gave rise to viral variants, such as European variants of HPV type 16 oncogene E6 (HPV16-E6V). In the last two decades, several studies investigated the relationship of SNPs in oncogene E6 and the differences in oncogenic potential of virus, rising E6 oncogene status to a potential therapeutic target. The aim of this work was to predict and evaluate, through the use of *in silico* methods, the differences between the binding sites of four European variants of HPV's oncogene E6. The binding site prediction was performed using the metaPocket 2.0 server. Based on the predicted coordinates of the binding site, we performed measurements of area and volume, along with electrostatic potential and hydrophobicity analysis of the studied region through the computational software USC CHIMERA. Our results showed that the binding site includes both zinc-binding domains and the α -helix that connects them. All variants displayed the same binding site region, but with meaningful differences in their respective area and volume. We did not find relevant differences between the electrostatic potential or hydrophobic profile of the binding site of different variants. To the present day, we have no knowledge of non-invasive therapeutic routines against HPV infection. The interaction between oncogene E6 with different tumor-suppressor proteins is directly related to the immortalization and uncontrolled growth of epithelial cells, which can lead to cervical cancer. By thoroughly uncovering the structural properties of the binding sites of E6, we can hopefully contribute to the future development of anti-HPV drugs.

Área: (3) Proteins and Proteomics

Analysis of Amino Acid coevolved sets in the Low Molecular Weight Phosphatase protein family by Molecular Dynamics

Marcelo Afonso, Lucas Bleicher

Universidade Federal de Minas Gerais

Low molecular weight phosphatases (LMW-PTPs) are one of the three existing major types of Tyrosine Phosphatases, with important roles in intracellular signalling of processes such as cellular growth, differentiation and proliferation through interactions with various possible substrates. By utilizing our group's technique for detecting amino acid correlation on the LMW-PTP protein family multiple sequence alignment and through bibliographic revision of articles that experiment or discuss these positions, we elucidated various amino acid coevolved sets and their possible biological meanings. Coevolved residue sets corresponding to the active enzymatic site and important active site hydrogen bonds were found, as well as coevolved sets that seem to be related to the active site cavity difference in charges between Low Molecular Weight Phosphatases and a class of Arsenate Reductases that has long arisen in a group of this protein family bacterial sequences. A Proline and Glycine coevolved set that seems related to important structural enzymatic properties has also been found as well as a fifth set including a P-loop Glycine (G14) and Cysteine (C17). This Cysteine functional importance has already been established as a residue responsible for protecting the enzyme against irreversible oxidation during redox stress by the formation of a disulphide bridge with the catalytic Cysteine (C12). In contrast there is no experimental or theoretical data relating possible roles of the Glycine. An interesting proposition is that the reason this correlation is observed may be related to this Glycine's possible influence on this disulphide bridge formation. To further elaborate on this hypothesis we here report our preliminary results of a two microsecond simulation of wild-type Human cytoplasmic protein tyrosine phosphatase of the A form and a two microsecond simulation of this protein's G14A mutation in order to access the possibility of this glycine conferring the needed P-loop flexibility for the formation of the catalytic C12 and C17 disulphide bridge.

In silico intrinsic disorder analysis of β -crystallin B2 protein and mutations that cause congenital cataract

J. B. O. Souza, J. E. A. Júnior, E. R. Tomarozzi and S. Giulietti

FMRP/USP

The congenital cataract is one of the major causes of visual deprivation in the world affecting 3 in every 10,000 live newborns in developing countries. Approximately 50% of children's cataracts are due to the genetic cause and the majority of autosomal dominant. The condition may be identified mainly based on the opacity in the ocular lens, and it is the result of some specific proteins loss of function. Singular mutations in the beta-crystalline B2 protein (CRYBB2) are one of the major causes of congenital cataract that affects several members of the same family, emphasizing the importance of this protein in the physiopathology of the disease. The CRYBB2 has regions of intrinsic disorder (ID) that can be indispensable in their role. ID regions show high number of interaction sites which are associated with cell signaling, becoming the target so that the unwanted interactions can be avoided. Thus, changes in the particular pattern of the ID proteins can cause cellular malfunction, favoring the occurrence of diseases. The aim of this study was to predict and analyze through in silico tools, the CRYBB2 protein intrinsic disorder (ID) and A2V, I21N, S31W, W59C, D128V, V146M, W151C, V187M and R188H mutant proteins, that cause congenital cataracts, in order to check for changes in the ID pattern caused by singular mutations in the protein. The ID's prediction was performed by ANCHOR and IUPred computational tools. The ANCHOR tool showed three ID regions. The first region is in interval of 15-25 aminoacids, where the I21N mutation occur that causes ID decrease in the mutated protein. The second one is in interval of 52-60, where the W59C mutation happen that caused ID increase. The third region is in interval of 151-16 where the W151C mutation is located which also causes increased ID. The IUPred tool also showed 3 ID regions. The first one in interval of aminoacids 1- 42, where the A2V, I21N and S3I mutations happen which cause ID decrease in this region. The second region is in interval of 94-99, however, to date, there are no known mutations in this range. The third region is in interval of 175-190, where the V187M and R188H mutations occur, where the V187M mutation causes an ID increase and R188H mutation did not change in the ID wild CRYBB2 protein pattern. Therefore, it is concluded that missense mutations were enough to promote alterations in the CRYBB2 mutated protein intrinsic disorder pattern.

PPI-signature: detecting similar interactions among homologous proteins and distinct partners

Larissa F. Leijôto, Raquel C. Melo-Minardi

*Universidade Federal de Minas Gerais
Departamento de Ciência da Computação*

About 80% of proteins are only capable of performing their functions through associations with other proteins. Different types of interactions are responsible for making such associations, and they are fundamental for stabilizing complexes and to ensure proteins will function properly. Furthermore, to optimize their functional roles, proteins interact with a spectrum of binding affinities, making these interactions the heart of most biological processes. Although there are a variety of affinities, Protein-Protein Interactions (PPIs) maintain a high degree of specificity for their partners. Thereby, the underlying premise is that there is a limited set of residues which participate in protein binding sites, and they are well-conserved to keep a specific way of interaction with other set of proteins.

Understanding how PPIs occur, and which are the specific interactions between a protein and their binding partner are crucial for explaining the structural and physicochemical determinants. Thus, we can shed light on how protein recognition and binding affinity takes effect. Moreover, this understanding can help in some applications. For instance, in proteins engineering, it can improve the design of resistant proteins to inhibition, and in prioritizing drug targets. Many works address their subject to discover a pattern in a protein family; however, few of them are concerned about the semantic of this pattern. The task of determining patterns in a complex is more complicated than in a monomer; thus, the existing methods are not good enough to identify patterns among interfaces. This failure occurs because most of the algorithms do not consider the whole protein complex, and the intrinsic features that a protein needs to bind in another one.

Finding a signature that describes which interactions are common in proteins, which belong to the same family, is vital. A signature can have significant implications for understanding the nature and function of PPIs, especially those that are considered to have "promiscuity." Hence, we proposed a multi-objective genetic algorithm to find patterns that are not straightforward. The algorithm proposed uses types of interactions that amino acids establish with their neighbors in three-dimensional space; also, it compares the residue identity and the environment associated with each of them. Consequently, we can point out what and how important the similar interactions among a set of interfaces are. The results have shown that this methodology is more appropriate to embrace characteristics that are relevant in interface comparisons, and on the identification of a functional signature.

Isocitrate Lyase of *Paracoccidioides brasiliensis*: Effects of Cofactor on Dynamical Stability and Virtual Screening of Natural Products

Luciane Sussuchi da Silva¹, Uessiley Ribeiro Barbosa¹, Fausto Guimarães Costa¹,
Célia Maria de Almeida Soares², Maristela Pereira² and Roosevelt Alves da Silva¹

Núcleo Colaborativo de Biossistemas, Federal University of Goiás¹; Departamento de
Bioquímica e Biologia Molecular, Federal University of Goiás²

The enzyme isocitrate lyase (ICL) catalyzes the cleavage of isocitrate into glyoxylate and succinate. ICL and the entire glyoxylate cycle are known to be involved in virulence and pathogenicity of human pathogenic bacteria and fungi. The absence of this enzyme in mammals makes it an interesting target for design of specific inhibitors, with more selectivity and fewer side effects. In this work, we aim the ICL of the dimorphic fungus *Paracoccidioides brasiliensis* (*PbICL*) for *in silico* searching and designing of new antifungal compounds. Magnesium ion have been seen required for full activity of *PbICL* and have been proposed to be involved on stabilization of the substrate during enzymatic catalysis in ICL superfamily. A homology model for *PbICL* was built with I-TASSER server based on ICL structure of *Aspergillus nidulans* (PDB 1DQU) and its magnesium-binding site was modeled based on the ICL structure of *Magnaporthe oryzae* (PDB 5E9F). The effects of the ion cofactor in the structural stability were evaluated through 100 ns of molecular dynamics simulation (MD) using 99SB-IDLN on GROMACS package. After cluster analysis, two ICL conformations from MD were selected for *in silico* virtual screening: one representative structure of the MD trajectory (cuto 0.3 nm) and another applying the isocitrate binding as a positive control among the cluster structures (cuto 0.15 nm). Virtual screening using AutoDock Vina were performed with 89399 natural products from ZINC database aiming at the cavity of the cofactor bind site. For each ICL conformation, 20 best ligands were selected taking into account criteria of affinity, efficiency and Tanimoto index. The same procedure of molecular dynamics and virtual screening was done for the enzyme in the absence of the magnesium ion. Overall structural dynamics, accessible surface area of the magnesium-binding site and topological profiles of the best ligand compounds were compared in the presence and absence of the cofactor.

Supported by: CNPq, CAPES, FINEP e FAPEG.

Peptides modulators of malate synthase of *Paracoccidioides brasiliensis* obtained from Protein-Protein interactions and docking simulations

Raisa Melo Lima, Luciane Sussuchi, Gabriela Lima de Menezes, Célia Maria de Almeida Soares, Maristela Pereira and Roosevelt Alves da Silva

Universidade Federal de Goiás

Paracoccidioidomycosis (PCM) is a systemic mycosis endemic in Brazil, where are recorded about 80% of cases worldwide, and It has *Paracoccidioides* sp. as the etiologic agent. Malate synthase of *Paracoccidioides* species (*PbMLS*) is an important enzyme related to the fungal metabolism, once it is essential in the glyoxylate cycle, a secondary metabolic pathway of the citric acid cycle exclusive to microorganisms and plants. Its absence in humans makes this enzyme an interesting subject to study, mainly in rational drug design. From recent in vitro studies, several interacting proteins of *PbMLS* were classified, but the modes of interaction and key regions involved in protein-protein interfaces (PPIs) have not been described yet. In this work, six (6) binding proteins (BPs) were selected to describe the PPI's of MLS. Their 3D structures, as well as *PbMLS*, were predicted by homology modeling using I-TASSER server, and subsequent molecular dynamics simulations (MD). The most common conformational modes of each protein were obtained by cluster analysis of the trajectories generated by MD. Molecular docking simulations using Gramm-X were then performed for the conformational modes of *PbMLS* against the BPs, resulting in a total of 36 complexes. Based on the higher frequency of some small fragments of proteins observed in the IPP's, 57 peptides with sizes between 5 and 20 residues, were initially selected from five regions of *PbMLS*, those considered more frequent in the protein-protein interactions. FlexPepDock simulations were performed to optimize the atomic coordinates of the peptide complexed with *PbMLS*, and concomitantly, PepFOLD simulations were performed to evaluate the stability of each peptide in solution. Based on the lower energy of peptides linked to *PbMLS*, as well as the stability of their structures in solution, six (6) peptides were selected as promising ligands to *PbMLS*. The stability and patterns of interactions of these peptides are showed in detail.

Supported by CNPq, FINEP, CAPES and FAPEG

Inhibition Resistance Mechanism for the Product of Beta-Glucosidases, a Computational Approach

Rafael E. O. Rocha, Leonardo H. F. de Lima

Federal University of Minas Gerais, Federal University of São João del Rei

Biofuels are a renewable energy source that are garnering global attention to the optimization and utilization of them. Such fuels have great relevance for decreasing dependence on fossil fuels, but may represent a obstacle to compete with the production of food from the sugar industry. Fortunately, the production of biofuels from by-products rich in cellulose of that industry, the second generation biofuels, show a potential to overcome that obstacle. They are produced from the use of a set of lignocellulitics microbial proteins. Such proteins degrade cellulose to fermentable sugars by an intricate chain of events. The last link in this chain is proving the biggest challenge in optimizing the production of biofuels, the conversion of cellobiose, coming from the previous steps, in free glucose. Most of the proteins responsible for this catalysis, the beta-glucosidases, are inhibited by the product, drastically reducing the yield of the process as a whole. However, GH1 beta-glucosidase family reports of resistance to inhibition may be observed, but such a mechanism is poorly understood. In this work, we try to understand the factors that influence the inhibition and resistance to inhibition of beta-glucosidases utilizing modeling and molecular dynamics tools. For this, we performed molecular dynamics of equilibrium simulation for two beta-glicosidases, a resistant protein (GH1) and a non-resistant protein (GH3) in the presence of cellobiose, docked glucoses and glucoses manually lysed, totalling six systems. These simulations were performed under CHARMM force field using the free software NAMD. Initially, the energy analysis suggests that the portion of the cellobiose close to catalytic triad (monomer -1) has an major affinity for the active site in GH3 than GH1, but this difference can not be observed when cellobiose is converted into glucose. Moreover, structural analysis seems to show that in GH1, cellobiose -1 monomer tends to be shifted close to sub-site +1 of the protein (more aromatic), reflection of the pi-stacking interactions between the +1 monomer and the same sub-site of the protein, thus losing several contacts between the monomer -1 and the sub-site -1 (more hydrophilic), but during the simulations with glucose (dock and lysis), the monomer -1 is freed to immediately interact with the sub-site -1, reflecting an increase in hydrogen interactions and decrease in hydrophobic contacts when compared to simulation with cellobiose and simulations with glucose. There were no significant energy differences among +1 glucoses in GH1 and GH3, suggesting that the bottleneck for catalysis is present in glucoses -1. Normal modes and principal components analysis are being made, seeking to find structural changes that might point to a possible opening backdoor in beta-glucosidase of the GH1 family. It is expected at the end of this work, we can contribute to a better understanding of the mechanism of inhibition and resistance to inhibition by product in beta-glucosidases.

Thanks for CAPES.

FlexSPS: A Monte Carlo update for protein refinement from I-TASSER models

Roosevelt Alves da Silva

Universidade Federal de Goiás

Comparative modeling is the main technique used in the protein structure prediction. 3D models are resolved based in the sequence alignment between the target and templates extracted from experimental protein structures in order to use the structural information of the aligned regions (from template) to construct a structural model for the target protein. Significant progress has been made in order to achieve high resolution models. The I-TASSER program has been extensively used by the community and it has been ranked as the best methodology for the protein structure prediction. However, this tool still requires improvements in order to increase the resolution of the predicted models, particularly avoiding the presence of steric clashes, unfavorable torsion angles, and unphysical bond length and bond angles. In this work, we have developed a tool for this purpose. The initial challenge was ensure improvement to the molProbity score and TMscore. Based on this initial goal, we have developed a tool called FlexSPS (Flexible Sampling for Protein Structure) that it allows sampling a large set of conformations around the predicted structure from I-Tasser. FlexSPS is a Monte Carlo program that modifies locally the protein conformations with the help of the *replica exchange method* around solutions from Amber forcefield and a ramachandran potential used to represent the protein. Our preliminary results indicate a substantial improvement in molProbity score and parameters related to the quality of the structures. However, changes in the force field will be needed to provide an increase in TMscore values. FlexSPS has been developed with the purpose of its application to the refinement of protein structures, protein-protein molecular docking, mutagenesis and loop predictions. The main application with this integrated tool will be the search for new drugs.

Supported by CNPq, FINEP, CAPES and FAPEG

Identification of non-homologous isofunctional enzymes in the antioxidant system of plants and phytopathogens

¹Rangeline Azevedo da Silva, ²Leandro de Mattos Pereira, ¹Melise Chaves da Silveira, ²Monete Rajão Gomes, ³Ana Carolina Guimarães, ¹Antonio Basílio de Miranda

¹FIOCRUZ - IOC – Laboratório de Biologia Computacional e Sistemas, ²Pontifícia Universidade Católica do Rio Grande do Sul, FIOCRUZ -IOC - ³Laboratório de Genômica Funcional e Bioinformática

In plants, the antioxidant system is responsible for the hypersensitive response, elicited when hosts are infected by pathogenic strains of bacteria or biotrophic fungi. It is composed of several enzymes such as catalase (CAT), peroxidase (POX), superoxide dismutase (SOD), Glutathione peroxidase (Gpx), Peroxiredoxin (Prxs), among others. This arsenal plays a critical role in the detoxification of reactive oxygen species during host-pathogen interactions. Blocking or inhibiting these enzymes would, in principle, decrease the virulence of the pathogen and/or delay the defense against free radicals used by the plant as defense mechanisms during the raids, debilitating the pathogen. Recently, researchers identified and characterized non-homologous isofunctional enzymes (NISEs), enzymes that perform the same biochemical function but have different evolutionary origins, which are reflected in differences between their primary and tertiary structures. These differences may be exploited for the development of specific blocking or inhibiting agents, resulting in a diminished virulence or pathogenicity. The objective of this study was to identify NISEs in the antioxidant system, using as model *Glycine max* and some of its pathogens, like *Aspergillus avus*, *Fusarium oxysporum*, *Phytophthora sojae*, *Sclerotinia sclerotiorum*, *Xanthomonas axonopodis*. We have also included *Apis mellifera* (pollinator), *Bacillus subtilis* (soil bacteria), *Azobacter chroococcum* and *Trichoderma arzianum* (soil fungi) and *Homo sapiens* in the analysis, for applications in a ecological context. Files containing information about active proteins in metabolic pathways were obtained from KEGG, and datasets of predicted proteins of *G. max* and its pathogens were downloaded from UniprotKB and RefSeq. The AnEnPi pipeline was used for clustering, by comparing the primary structures of enzymes previously annotated with the same Enzyme Commission number. The BLASTP was used to analyze the difference between the primary structures of the enzymes within each EC, on this step the activities were compared all against all. After clustering, enzymes grouped in the same cluster were considered homologous (score above 120), while enzymes allocated in different clusters were considered potential analogous enzymes (score under 120). The identified NISEs had their folds sorted using the SCOP and SUPERFAMILY databases. In this work, we have been able to identify several NISEs candidates belonging to the antioxidant system between *G. max* and its pathogens: 9 for CAT, 7 for POX, 6 for SOD and 1 Prxs. These results show that it may be possible to exploit differences in the enzymes belonging to the antioxidant system to develop specific inhibitor molecules.

Funding support: CAPES - Coordenação de Aperfeiçoamento de Pessoal de Nível Superior.

In silico study of Hypoxanthine-guanine Phosphoribosyltransferase inhibitors for drug design against Leishmania species

L.P. Araújo^{1,5}, W. R. A. Soares^{2,5}, R.S. Pereira^{3,5}, B. S. Andrade^{4,5}

¹Departamento de Química e Exatas, Universidade Estadual do Sudoeste da Bahia (UESB); ²Departamento de Saúde II, UESB; ³Programa de Pós-graduação em Química,

UESB, Brazil; ⁴Departamento de Ciências Biológicas, UESB; ⁵Laboratório de

Bioinformática e Química Computacional – LBQC/UESB, Campus Jequié, Brazil.

The hypoxanthine-guanine phosphoribosyltransferase (HPRT) initiates the metabolism of toxic purine bases of the *Leishmania* species. This enzyme mechanism is specific for the parasite and is absent in mammalian hosts. This factor allows the HPRT becomes a promising target for drug development on leishmaniasis treatment. The active site of HPRT presents one guanine monophosphate (GMP) in both chains A and B. The main difference between *Leishmania* and human enzymes is the interaction between the ribose GMP, which allows a comparison between these enzymes and contributes to the exploration of potential inhibitors. The ligands used in this work are deposited in The SAM Database, hosted at the Laboratory of Bioinformatics and Computational Chemistry from Universidade Estadual do Sudoeste da Bahia, Jequié, Brazil. All compounds were isolated from Brazilian semiarid plant species, with active extracts against *Leishmania* species. The molecular structures of the ligands were prepared in Marvin Sketch (Chemaxon) for verification of the valences and structural errors, and then saved in mol2 format. The AutoDock tools was used to prepares the ligands for docking studies and convert them to pdbqt format. Furthermore, it was defined the region of interaction with potential inhibitors (gridbox) and recorded coordinates. The molecular docking calculation was performed using the AutoDock Vina program. Ligand docking poses with better energy values were choosed using PyMOL 1.7, as well as for obtaining the complexes in pdb format. The interaction maps for each best energy complex were generated using Discovery Studio 4.0 program. We selected 82 deposited compound structures, isolated from Brazilian semiarid plants, in order to perform virtual screening and molecular docking studies for HPRT inhibitors, based on previous studies of these plant extracts against *Leishmania* species. Seven structures presented affinity energies below -7.0 Kcal/Mol, and all ligands presenting interactions with active site residues of HPRT. The molecule SAM25442 presented the best affinity energy of -7.8 Kcal/Mol. The next steps of this study are generating analogues for each best affinity compound found, as well as free-energy calculations using molecular dynamics approaches. These chemical constituents can become future drug candidates against *Leishmania* species.

In silico screening of semiarid plant compounds targeting 5-lipoxygenase (LOX)

L.P. Araújo^{1,5}, B.S. Portela^{1,5}, W. R. A. Soares^{2,5}, R.S. Pereira^{3,5}, B. S. Andrade^{4,5}

¹Departamento de Química e Exatas, Universidade Estadual do Sudoeste da Bahia (UESB);

²Departamento de Saúde II, UESB; ³Programa de Pós-graduação em Química,

UESB, Brazil; ⁴Departamento de Ciências Biológicas, UESB; ⁵Laboratório de

Bioinformática e Química Computacional – LBQC/UESB, Campus Jequié, Brazil

5-Lipoxygenase (LOX) are a family of enzymes important in the production of essential chemical mediators in the inflammatory process and allergies. This enzyme is a therapeutic target used in the production of medicaments leukotriene antagonists (LTA) versus obstructive respiratory diseases (asthma, bronchitis, allergic rhinitis). The ligands used in this work are deposited in The SAM Database, hosted at the Laboratory of Bioinformatics and Computational Chemistry from Universidade Estadual do Sudoeste da Bahia, Jequié, Brazil. All compounds were isolated from brazilian semiarid plant species, with active extracts with previous described antiinflammatory potential. Molecular structures of all ligands were prepared in Marvin Sketch (Chemaxon) for verification of the valences and structural errors, and then saved in Mol2 format. AutoDock tools was used for ligand preparation for docking studies and convert them to pdbqt format. Furthermore, it was defined the region of interaction with potential inhibitors (gridbox) and recorded coordinates, based on LOX crystallographic active site described on the literature. The molecular docking calculation was performed using the AutoDock Vina. Ligand docking poses with better energy values were chosen using PyMOL 1.7, and then the complexes were saved in pdb format. The protein-ligand interaction maps for each best energy complex were generated using Discovery Studio 4.0, in order to verify if each ligand interacted with the active pocket residues. We selected 26 deposited compound structures, isolated from brazilian semiarid plants, in order to perform virtual screening and molecular docking studies against LOX. After Autodock Vina calculations and pose validations, 16 structures presented affinity energies bellow -7.0 Kcal/Mol interacting with LOX active site residues. The molecule SAM2725 presented the best affinity energy (-9.9 Kcal/Mol) and the best positioning in the pocket (PHE177 and GLN363), in comparison to montelukast® (-8.6 Kcal/Mol), an inhibitor of leukotriene synthesis. Even virtual screening approach using Autodock Vina calculations has been validated for 190 protein-ligand complexes bellow 2.0 Å, is very important consider free-energy molecular dynamics calculations with solvent accessibility (eg. MMPBS/GBSA) to confirm SAM2725 as a theoretical LOX inhibitor. In addition to molecular dynamics, in a further step we will generate analogues for each best affinity compound found. As a preliminary screening, this study may provide chemical natural compounds to be tested in vitro as new inhibitors of LOX for future drug candidates.

Virtual screening of natural compounds from Brazilian semiarid plants targeting GABA receptor inhibitors

W. R. A. Soares, G. Santos, D.M. Oliveira, V.F. De Paula, B. S. Andrade

Departamento de Saúde II. Universidade Estadual do Sudoeste da Bahia, Campus Jequié, Brazil; Departamento de Química e Exatas. Universidade Estadual do Sudoeste da Bahia, Campus Jequié, Brazil; Departamento de Ciências Biológicas. Universidade Estadual do Sudoeste da Bahia, Campus Jequié, Brazil; Laboratório de Bioinformática e Química Computacional – LBQC/UESB. Universidade Estadual do Sudoeste da Bahia, Campus Jequié, Brazil.

Different Brazilian semiarid plant species are widely used in popular medicine because they have sedative, anxiolytic, anticonvulsant and central analgesic effects. This study investigated *in silico* psychopharmacological action of these plant compounds by docking them human GABAA receptor. The ligands used in this work are deposited in The SAM Database, hosted at the Laboratory of Bioinformatics and Computational Chemistry from Universidade Estadual do Sudoeste da Bahia, Jequié, Brazil. All compounds were isolated from Brazilian semiarid plant species with psychopharmacological activity. The molecular structures of the ligands were prepared in Marvin Sketch (Chemaxon) for verification of the valences and structural errors, and then saved in mol2 format. The AutoDock tools was used to prepare the ligands for docking studies and convert them to pdbqt format. Furthermore, it was defined the region of interaction with potential inhibitors (gridbox) and recorded coordinates. The molecular docking calculation was performed using the AutoDock Vina program. Ligand docking poses with better energy values were choosed using PyMOL 1.7, as well as for obtaining the complexes in pdb format. The interaction maps for each best energy complex were generated using Discovery Studio 4.0 program. We tested 78 chemical structures in molecular docking studies described for Brazilian semiarid plants against GABA receptor. Even different structures have presented good interactions with the receptor, SAM2800 and SAM 3201 compounds showed the highest interaction with the target, and presented better affinity energies (-10,2 kcal/Mol) in comparison to Diazepam® (-9,0 kcal/Mol). Therefore, this study reports the need for further research on the extracts and isolated compounds *in vitro* and *in vivo* in order to validate the therapeutic properties of these plants, and thus its chemical constituents can become future drug candidates.

24-c-sterol-methyltransferase as a target for the design of new anti- trypanosomatids drugs

Kendy Anny de Azevedo Werneck, Gonzalo Guillermo Visbal Silva, Diego Enry
Barreto Gomes, Manuela Leal da Silva

*Diretoria de Metrologia Aplicada às Ciências da Vida - Dimav, Instituto Nacional de
Metrologia, Qualidade e Tecnologia - Inmetro, Duque de Caxias/RJ, Brazil*

Trypanosoma brucei is the etiologic agent of sleeping sickness, a neglected tropical disease (NTD) affecting mostly low-income populations in tropical countries. The last product of sterol biosynthesis in parasitic is ergosterol and 24-alkylated sterols are major cell membrane components of parasites, in contrast to cholesterol in mammals. Their biosynthesis requires an alkylation, catalyzed by an S-adenosyl-L-methionine: Δ24- sterol methyltransferase (Tb24-SMT), a key difference between cholesterol and ergosterol biosynthesis, presenting an opportunity for the rational design of anti- infective agents. Its inhibition prevents the survival of parasites and is an important target for design of anti-trypanosomatids drugs. However, a limiting factor for the structural-based drug design is the absence of a Tb24-SMT 3D model for an appropriate active site mapping and pharmacophore models. In this work, we built 100 3D models candidate for Tb24-SMT, using comparative modeling and threading approach, based on the structure of 4'-O-methyltransferase from *L. aerocolonigenes* (PDBid 3BUS) with 60% coverage, 23% identity and 41% similarity between sequences. The best model was chosen and validated through the parameters: Ramachandran plot where the highest value of R1 (residue in most favored regions) was 95.2%, the average deviation between template and model (RMSD) was 0.37Å. The active site residues being mapped as N134, Q139, D162, F163, M163, M166, I179. Molecular docking studies were performed with derivatives of azasterol inhibitors by AutoDock and Vina softwares. We conclude that sequential analyses identified the binding site residues and showed it is conserved between template and model. Molecular docking studies for all inhibitors showed more than 32% of conformations in the same cluster. The azasterol+3C (-12.90 kcal/mol), 22-piperidin-3yl-pregn-22(S),3b-diol (-12.43 kcal/mol) and 24-b-aminolanosterol (-12.06 kcal/mol) inhibitors demonstrated the best results with lower docking energy, better than the natural substrate zimosterol (-10.10 kcal/mol) and lanosterol (-10.40 kcal/mol) results for this enzyme. The amino acids participating in the interaction between enzyme and compounds were mapped. Next step consists in MD simulations for providing detailed information on the interaction and fluctuations in each complex with compounds for further experimental purposes.

Financial support: CAPES and CNPq

Isocitrate lyase protein-protein interaction assay of *Paracoccidioides* spp.

Kleber Santiago Freitas e Silva, Célia Maria de Almeida Soares, Maristela Pereira

Universidade Federal de Goiás, Goiânia, Goiás

The fungus *Paracoccidioides* spp. is the causative agent of paracoccidioidomycosis (PCM) which is a pulmonary fungal infection. The disease develops after inhalation of fungal propagules that reach the alveolar epithelium in the lungs and then they differentiate into the pathogenic yeast form. The isocitrate lyase (ICL) is a key enzyme for the glyoxylate cycle, it is present in the fungi but absent in human. Thus, ICL is an important target in the pursuit of inhibitors, since it would present little or no toxicity to humans. In fungi, it has been shown that ICL and the entire glyoxylate pathway enzymes are generally induced under conditions of low glucose and low oxygen tension and especially in the presence of acetate. In *Paracoccidioides* sp., our group has shown that the *Paracoccidioides* sp. isocitrate lyase (*PbICL*) transcript and protein levels are the same for glucose and acetate. However, the ICL activity is higher in acetate than in glucose, being regulated by phosphorylation. The objective of this study is to identify and analyze the fungus proteins that are likely to bind to *PbICL*. It is well-known that protein interactions are intrinsic to cell processes, and it may be possible to infer the function of a protein through the identification of its ligands. Yeast, transition and mycelium protein crude extracts were obtained by disruption of cells in the presence of protease inhibitors. The mixture was centrifuged and the supernatant was used for further analysis of proteins by one-dimensional gel electrophoresis. Yeast cells were grown for 7 days in solid medium and mycelium was grown for 15 days also in solid medium. The purified recombinant *PbICL* was used to produce anti-*PbICL* polyclonal serum in mice. The investigation for interactions is performed through an in vitro assay. We identified more than 600 proteins that bind to *PbICL* in the yeast phase and more than 150 proteins that bind to ICL during mycelium and transition phases. We compared this results to the ones published in the scientific literature through String database and we realized that most of the proteins that bind to *PbICL* have not been identified before. In silico and docking analysis allied to virtual screening will be essential in order to continue the investigation of *PbICL* inhibitors, leading to the finding of potential antifungal with minimal side effects.

Financial Support: Capes

In silico repurposing of approved drugs for paracoccidioidomycosis

Amanda Alves de Oliveira¹, Bruno Junior Neves², Kleber Santiago Freitas e Silva¹, Lívia do Carmo Silva¹, Célia Maria de Almeida Soares¹, Carolina H. Andrade², Maristela Pereira¹

¹Laboratório de Biologia Molecular, Instituto de Ciências Biológicas, Universidade Federal de Goiás, Goiânia, Brasil; ²Laboratório de Planejamento de Fármacos e Modelagem Molecular, Faculdade de Farmácia, Universidade Federal de Goiás, Goiânia, Brasil

Paracoccidioides spp. is a thermodimorphic fungus associated with paracoccidioidomycosis (PCM), the most common systemic mycosis in Latin America. The PCM treatment involves a long-term chemotherapeutic approach and relapses occur at an alarming frequency. Moreover, the emergence of strains with increased drug-resistance phenotypes makes the discovery of new drugs an urgent task. Aiming at repurposing drugs for treating PCM, our group implemented an *in silico* chemogenomics screen on *Paracoccidioides* spp. genomes based on concept that "proteins sharing enough similarity (orthology) have enhanced the probability of share the same ligands". Initially, using the OrthoVenn web platform, we compiled a list of 6743 *P. brasiliensis* proteins (Pb01) with orthologous ($E\text{-value} \leq 10\text{--}20$) in other two isolates (Pb03 and Pb18). Then, protein sequences of the prioritized proteins were aligned against the sequence of drug targets in the DrugBank and TTD databases to screen for drugs that can potentially have anti-PCM activity. Inclusion and exclusion criteria such as drugs approved in phase I of the clinical studies, FDA approved status, sequence identity ($\geq 30\%$), and conservation state of functional amino acid residues were also incorporated in the drug screening. As a result, 254 proteins genes encoding potential *Paracoccidioides* drug targets for a total of 982 approved drugs or drug candidates were identified. Among the combined list of potential drugs targeting PCM proteins, drug such as Sulfamethoxazole, ketoconazole, Itraconazole, Fluconazole, Voriconazole, Rifampicin, already are used in PCM treatment. In doing so, we suggested an array of drugs that are expected to inhibit several metabolic processes, such as protein translation, transport $\text{Na}^+ - \text{K}^+ - 2\text{Cl}^-$, non-depolarising muscle relaxant and imprisonment cell cycle. The drugs identified were distributed within a wide range of classes, including antibiotics (e.g., Sulfoxone), anti-inflammatory agents (e.g. Acetylsalicylic acid), hypertensive agents (e.g., Chlorthalidone), antiepileptic agents (e.g. Vigabatrin,) and anticancer agents (e.g., Bosutinib). Our next step is to screen experimentally these drugs against Pb 01, Pb 03, and Pb18 and develop homology models and molecular docking studies to reveal insights into the molecular basis of drug action.

Financial Support: Capes, MCTI, CNPq, FNDCT, PRO-CENTRO-OESTE, and FAPEG

Screening of compounds candidate to inhibit the interaction monomer - monomer of the NS1 protein of dengue virus: an approach for docking and molecular dynamics.

Gonçalves R.L, Sussuchi L.A, Da Silva R.A

University Feral of Goiás

The disease caused by Dengue virus is a major public health problem in Brazil and several other developing tropical countries, affecting populations with low socioeconomic levels, making it a neglected disease. Dengue virus has encoded in its genome three structural proteins and seven nonstructural. The NS1 (nonstructural protein 1) is found in different oligomeric forms glycosylated and its functions are assigned only to the "mature" forms of the protein (dimer and hexamer with sugars). Therefore, inhibition of the maturation of this protein has been appointed as a good investment for the rational drug design. In this study, the partial structure of the monomer protein of NS1 (ID 4o6b) was used to generate the model through the Swiss Model Server. The structure with its 6 disulphide bonds and two sugars N-linked was submitted to 5 simulations of Molecular Dynamics (MD) for 30 nanoseconds (different seeds) using Gromacs software. The flexibility and structural stability of their domains were evaluated using all trajectories of simulations in order to select the conformations that best represent the structure around its native state. Compounds experimentally validated for homologous structures were used as positive control to assist in the selection of conformations that could result in high-affinity interaction with it. The conformations of NS1 with better stability had their target sites defined to be confronted with the ZINC database (89.415 structures of the natural compounds). The compounds were subjected to molecular docking with the program AutoDock Vina to select the best 100 compounds. In the next step, all these selected compounds were simulated again against the target, running 1000 independent simulations for each, to define its energy profile. Finally the best 10 compounds were selected and had evidenced their modes of interaction. The stability of these compounds against NS1 were also checked by MD simulations, where the stability of the interaction showed the best binders / inhibitors against the targets of NS1.

Supported by: CNPq, CAPES, Finep, FAPEG

Development of a peptide-based electrochemical biosensor for juvenile idiopathic arthritis diagnosis

Vinícius de Rezende Rodovalho¹, Galber Rodrigues Araujo², Carlos Ueira Vieira²,
João Marcos Madurro³, Ana Graci Brito-Madurro²

1. Institute of Biological Sciences, Federal University of Minas Gerais, MG, Brazil. 2. Institute of Genetics and Biochemistry, Federal University of Uberlândia, MG, Brazil. 3. Institute of Chemistry, Federal University of Uberlândia, MG, Brazil.

Juvenile idiopathic arthritis (JIA) is a wide group of autoimmune and inflammatory diseases that affect children and adolescents under the age of sixteen. In the absence of proper diagnosis and treatment, irreversible damage may occur in the joint tissues. Biosensors emerge in the clinical scenario as promising analytical alternatives to molecular diagnosis. In previous work, PRF+1 peptide (ACSSWLPRLPRGCAGGS) has been selected from random phage-display libraries and shown to react against antibodies from patients with JIA. This work presents the development of an electrochemical biosensor for the diagnosis of JIA in human serum using screen-printed carbon electrodes (SPCE) functionalized with PRF+1 peptide. Electrochemical methods and/or structural analysis were employed to investigate PRF+1 immobilization and its interaction with the antibody target through the monitoring of an oxidation peak. The SPCE surface was cleaned by cyclic voltammetry in perchloric acid (0.5 mol.L^{-1}) prior to immobilization of the peptide. Electrochemical detections were conducted by differential pulse voltammetry in phosphate buffer (0.1 mol.L^{-1} , pH 7.4). Moreover, PRF+1 three-dimensional structure was predicted with I-Tasser suite. Cyclic voltammograms in phosphate buffer showed a well-defined oxidation peak in $E_p = +0.49 \text{ V}$ for the electrode functionalized with the peptide (SPCE/PRF+1) and no peaks in this potential range for bare SPCE. Among the common 20 amino acids that compose peptides and proteins, methionine, tyrosine, histidine, tryptophan and cysteine are oxidized on carbon electrodes. However, only cysteine and tryptophan residues are present in PRF+1 sequence. Structural analysis was employed to investigate which of them was related with the detected oxidation peak. As the two cysteine residues form a disulfide bond, which is the oxidized form of these amino acids, the oxidation peak is probably due to tryptophan residue oxidation. This peak was also monitored for this system after blocking with bovine serum albumin (BSA) and incubation with positive or negative serum for JIA. The biosensor was able to discriminate samples from these different groups of patients. Therefore, a simple, miniaturized and functional platform was developed as a promising strategy for JIA molecular diagnosis.

Funding support: FAPEMIG, CAPES, CNPq.

Automation of polyproteins GAG and GAG-POL-1 cleavage site mapping: A study of large scale molecular docking

Fernando Limoeiro Lara de Oliveira , Maria Fernanda Ribeiro Dias. , Manuela Leal da Silva

Universidade Federal do Rio de Janeiro; Instituto Nacional de Metrologia, Qualidade e Tecnologia – Diretório de Metrologia Aplicada às Ciências da Vida (DIMAV)

Viral infections affect populations over the globe. The most of the treatments available for the population is based on prevention of infection rather than extermination of the virus. The difficulty for the design and development of antiviral drugs results from different viruses and their mutational capacity. The drugs in general acts on biomolecular mechanisms common to both viruses and hosts, making specificity for drugs a challenge, which will culminate in losses on health and life quality. In this context, the use of bioinformatics tools may provide a time and cost efficient approach for the study and development of antivirals, since such techniques can deal with the screening of huge amounts of biological and experimental data very rapidly. In this work, we are showing the early stages of the development of large-scale workflow for docking and mapping of the Gag and Gag-Pol polypeptides cleaved by the HIV-1 Protease. This is a well-studied mechanism of the HIV, in which GAG and GAG-pol are cleaved into active, smaller proteins. These polypeptides will act on the maturation of the HIV virion. Our goal is to map the cleavage sites of the polypeptides based on a more conformational approach, using the molecular docking. The first step of the workflow consists on the fragmentation of both gag and gag-pol into smaller peptides of different lengths (4, 6 and 8 amino acids). Each fragment size characterizes a different assay. We build a tridimensional structure of each peptide using the software Modeller, for both modeling and structure optimization. Each group of structures containing is automatically prepared for the molecular docking procedures using the OpenBabel and Autodock Tools softwares for a series of operations. Our preliminary tests on the workflow are being made on groups of fragments containing four amino acids. The peptides sequences were created using a Python script. The construction of the models resulted in satisfactory scores by Modeller, yielding a final number of 1445 structures, docking-ready. The next step is to perform the docking (AutodockVina) on the 4-residues GAG and GAG-pol peptide groups, in order to assess the optimal parameters for the mapping process. References and previous studies showed that peptide size, location and composition are crucial to determine substrate recognition by the enzyme. Therefore we believe that such methods, applied on natural substrates of enzymes, may provide useful information for the development of new maturation inhibitory drugs.

Financial support: CAPES and CNPq.

Characterization of the proteome of four strains of *Lactococcus lactis* with biotechnological relevance

Caroline Leonel Vasconcelos de Campos¹, Wanderson Marques da Silva¹, Cassiana Severiano de Sousa¹, Siomar de Castro Soares², Guilherme Campos³, Cristiana Perdigão Resende³, Felipe Luis Pereira³, Gustavo Henrique Souza⁴, Henrique Cesar Pereira Figueiredo³, Vasco Ariston de Carvalho Azevedo¹

¹Laboratório de Genética Celular e Molecular, Instituto de Ciências Biológicas, UFMG, Belo Horizonte, MG, Brasil, ²Universidade Federal do Triângulo Mineiro, Instituto de Ciências Biológicas e Naturais, Uberaba, MG – Brasil, ³Universidade Federal de Minas Gerais, Escola de Veterinária, UFMG, Belo Horizonte, MG, Brasil, ⁴Waters Corporation, Brasil.

Lactococcus lactis is one of the utmost studied organisms considering lactic acid bacteria. Due to its role on both human and animal diet, these bacteria gained prominence in the dairy industry. Moreover, *L. lactis* can be used as probiotic or as vehicles for the heterologous expression of molecules of interest in the human body. So, *L. lactis* is of great importance for the pharmaceutical industry. Given this biotechnological potential, a great variety of structural and functional studies of this bacterium genome was performed. However, no study to date has been performed in order to determine the core-proteome of *L. lactis*. Know the core-proteome of *L. lactis* will provide valuable information about a set of conserved proteins that may play vital physiological functions in the adaptive process of this specie and, consequently, contribute to biotechnological optimization. Thus, in order to expand our knowledge of the physiological molecular basis of *L. lactis* and to complement previous structural and functional studies of the genome of this bacterium, the present work used high-throughput proteomics to characterize the proteome of four *L. lactis* strains with great biotechnological relevance. NCDO2118, IL1403, MG1363 and NZ9000 were grown in primary medium of synthetic cultivation for such bacteria. The four strains were grown in M17 medium supplemented with 0,5% glucose, at 30°C, for 16 hours, with no shaking. After obtaining proteins from total bacterial lysate, they were digested with trypsin. The tryptic fragments generated were subjected to proteomic analysis by LC / MS. From the proteome of the four strains, a total of 1109 *L. lactis* non-redundant protein were identified. Comparing this result with the in silico data of *L. lactis* core-genome, it was able to validate 56% (946 proteins) of the ORF encoding the predicted core genome. According to the analysis of Blast2Go tool (predicting biological processes), the proteins that comprise the core proteome were grouped into 20 biological processes. The processes with the highest number of proteins were: translation (96 proteins), amino acid metabolism (37 proteins), nucleotide metabolism (34 proteins) and carbohydrate metabolism (32 proteins). The core proteome was also analysed by KEEG database to assess which metabolic pathways were active and enrichment analysis showed the following processes: Ribosome ($P = 0.0110$), carbon metabolism ($P = 0.0336$) and pyruvate metabolism ($P = 0.04280$). The results obtained to date demonstrated the key proteins and the metabolic processes that contribute to the growth of *L. lactis* in M17.

Refining the calibration of a coarse-grained force field for protein complexation

Sergio Alejandro Poveda Cuevas¹ and Fernando Luís Barroso da Silva^{1,2}

¹Bioinfo-USP/Brazil, ²DFQ/FCFRP-USP/Brazil

Theoretical studies of the molecular mechanisms responsible for the formation and stability of protein complexes have gained importance due to their practical applications in the understanding of the molecular basis of several diseases, in protein engineering and biotechnology. The objective of this work is to refine a constant-pH coarse-grained force field for protein-protein interactions based on experimental thermodynamic parameter called second virial coefficient. Our ultimate goal is to generate knowledge for a better understanding of the physical mechanisms responsible for the protein associations in different environments. By means of computational tools and mesoscopic models solved by Monte Carlo simulations, the homo-association of lysozyme and chymotrypsinogen was used in the calibration process. Our model is composed of two contributions namely Coulombic (with charge fluctuation as a function of the pH) and Lennard-Jones. A better description for van der Waals contribution was necessary to improve the quantitative theoretical results in different salt regimes. Our main interest is focused on how to deal with hydrophobic molecules within the continuum dielectric framework. We found that there is an inherent difficulty in reproducing the experimental data when low salt concentrations and pH values far from pI are considered. However, considering moderate salt concentrations a reliable description of the experimental data is possible. There is an ongoing study to discuss how the exposition of hydrophobic amino acids at pH conditions far from pI can be still captured in this simplified model.

Supported by: CNPq and Fapesp.

Preliminary analysis of microRNAs and their pathway genes in the genome of *Globodera pallida*.

Carlos Bruno de Araújo¹; Caio Borges Melo¹; Julia Silveira Queiroz¹; Laurence Rodrigues do Amaral¹; Matheus de Souza Gomes¹.

¹Laboratory of Bioinformatics and Molecular Analysis, Institute of Genetic and Biochemistry/ Faculty of Computing, Federal University of Uberlândia, Campus Patos de Minas, MG, Brazil.

The *Globodera pallida* or potato cyst nematode (PCN) is one of the more specialized species of plant-parasitic nematode. *G. pallida* is a serious pest of potatoes around the world and causes loss of approximately 50% of European production. One of the most representative small noncoding RNA class is the microRNAs (miRNAs). They're responsible for gene expression control posttranscriptionally involved in multiples biological processes. This control is performed by transcriptional gene silencing, inhibition of translation or RNA-target cleavage. The miRNA processing pathway has several proteins including the key proteins of the pathway Argonaute, Dicer and Drosha. Recently, the genome of *G. pallida* was sequenced and deposited in WormBase – ParaSite. The aim of this work was to identify and characterize the miRNAs and miRNA pathway genes in the genome of *G. pallida* using bioinformatics approaches. Proteins from *Drosophila melanogaster* and *Caenorhabditis elegans* species were used as queries to search and predict the putative *G. pallida* miRNA pathway proteins. We also characterized the distribution of the conserved domains and also conserved amino acids in the active sites. The putative Argonaute and Dicer proteins predicted in *G. pallida* displayed highly conserved domains and amino acid in active sites when compared to their orthologue proteins. Besides phylogenetic analysis confirmed the conservation of the *G. pallida* proteins with their respective orthologues. Using a robust algorithm, we also found 51 mature miRNAs and 68 precursor miRNAs in the genome of *G. pallida*. We analyzed the thermodynamic and structural features of the precursor miRNAs and the conservation of the secondary and primary structure of the miRNAs identified. We found that several *G. pallida* miRNAs were much conserved compared to miRNAs from different species. The miRNAs glo-miR-87 and glo-miR-124 were 100% conserved when compared to miRNAs from nematode species displaying high conservation at mature miRNA level. The GC content and MFE (minimal free energy) values of *G. pallida* miRNA precursors displayed high conservation compared to ortholog miRNA values. The finding of this work provided better knowledge about miRNAs and miRNA pathway genes in nematode species collaborating with the understanding of the essential biological processes in *G. pallida*. The data obtained are unprecedented and could help to elucidate new methods to control this pest in potato species.

Funding support: FAPEMIG

Human riboswitch: are we close to predicting it?

Deborah Antunes, Fabio Passetti, Ernesto Raul Caffarena

Fundação Oswaldo Cruz

RNA molecules are essential cellular players for many fundamental biological processes. Distinct RNA classes present structural features with specific functional roles in prokaryotes and eukaryotes. Similar to proteins, the RNA structure may be subject to changes due to the interaction with various ligands, including proteins, other RNAs, and metabolites. Riboswitch is a molecular mechanism in which the RNA structure changes based on specific metabolite-RNA binding. A riboswitch can regulate gene expression in different aspects, such as the attenuation of transcription, translation initiation, mRNA splicing and mRNA processing. In many cases, they are involved in the regulation of key genes, which makes its control an essential part of cell survival. Most studies on riboswitches investigated their existence in prokaryotic organisms. In eukaryotes, TPP riboswitches have been found in fungi, algae, and plants. In animals, riboswitches have yet to be identified. The focus of this work is to determine potential riboswitches in the human genome. The study was conducted in three stages: (i) search for candidate riboswitch in the human genome, using the Infernal software that searches DNA sequence databases for RNA secondary structure and sequence similarities; (ii) modeling the three-dimensional structure of a candidate sequence using ModeRNA software; (iii) molecular dynamics simulations using package Gromacs to verify the stability of the modeled structure and compare it to the template. A sequence candidate TPP riboswitch in the mRNA transcript variant FBLN2 gene was identified. This sequence has 55% identity with the *Escherichia coli* TPP riboswitch (PDB Id: 2GDI:X) and 84% conserved residues for this type of riboswitch. As the TPP riboswitch has its conserved 3D structure, we performed comparative modeling using the structure 2GDI:X as a template. The model presented RMSD of 0.7 Å with the template. Model and template were submitted to 1 μs of molecular dynamics simulations each. The template presented RMSD values of 0.36±0.02 while the model showed a higher value of 0.50±0.03 nm. Despite the conformational change, the model kept the base pairing and secondary structure without modifications in helix P1, P2, P5 and loop L5. Perspectives foresee a study on the candidate's affinity for the ligand, by molecular docking assays and new molecular dynamics simulations of RNA-ligand complex.

Financial support: Fiocruz, CAPES, FAPERJ, CNPq

miRNAs Expression and *in silico* Prediction of Targets Related with Resistant Exercise and Carbohydrate/Protein Supplementation

Souza, A.V.; Diaz, M.M.; Bocanegra, O.L.; Teixeira, R.R.; Siqueira, M.C.; Gomes, M.S.; Espindola, F.S.

Federal University of Uberlandia, Institute of Genetics and Biochemistry

MicroRNAs (miRNAs) are small non-coding molecules of RNA that regulates gene expression at the posttranscriptional level. To date, only a handful of studies have investigated changes on the levels of circulating miRNAs (c-miRNAs) in response to exercise. Here, we investigated the response of twelve c-miRNAs to resistance exercise (RE) and carbohydrate or carbohydrate/protein supplementation and evaluated the putative c-miRNA targets by using bioinformatics tool. Samples of blood were collected from 12 recreationally active young males before exercise, 03 and 24 hours afterwards. RT-qPCR was used for quantification of microRNAs and the relative expression data were analyzed using a two-way ANOVA with repeated measures. *In silico* prediction of miRNAs targets was performed using TargetScan version 7.1 (available at: <http://www.targetscan.org/>). As result, we observed four c-miRNAs with significant variation following RE: hsa-miR-133a, hsa-miR-503, hsa-miR-16 and hsa-miR-126. Acute ingestion of dietary protein elicited increases in the expression of hsa-miR-133 and -503 during the first 03 h after RE with decreasing levels over the following 24 h. Protein ingestion also led to decreased levels of hsa-miR-16 24 h after RE. Carbohydrate supplementation triggered increases on the levels of hsa-miR-126 shortly after RE. Among the putative c-miRNA targets, we found genes that have been implicated as regulators of cell proliferation, differentiation, and transformation (e.g. hsa-miR-133a target: FOSL2 gene - FOS-like antigen 2); cell cycle regulators (e.g. hsa-miR-503 targets: CCND2 gene - cyclin D2 and CDCA4 gene- cell division cycle associated 4); and angiogenesis (hsa-miR-16 target: C1QL gene- complement component 1; and hsa-miR-126 target: ADAM9 gene - ADAM metallopeptidase domain 9). Overall, our findings suggest a distinct profile of expression in c-miRNA between dietary carbohydrate and carbohydrate/protein supplementation following RE. As expected, the molecular response in the group that supplemented with protein was more pronounced for c-miRNAs involved in the regulation of myogenesis. Meantime, both treatments revealed a differential expression of c-miRNAs involved in angiogenesis. We conclude that *in silico* analysis can contribute to better understand the molecular mechanisms related to exercise adaptation mediated by miRNAs. Therefore, the differential expression of c-miRNAs and the effects on target genes predicted might be partially responsible for muscle hypertrophy and neovascularization following exercise.

Funding support: CNPq, FAPEMIG and PROPP/UFU.

Preliminary analysis of Dicer-like genes in *Cucumber* genome

Julia Silveira Queiroz¹; Tamires Caixeta Alves¹; Núbia Carolina Pereira Silva¹; Laurence Rodrigues do Amaral¹; Matheus de Souza Gomes¹

¹Laboratory of Bioinformatics and Molecular Analysis, Institute of Genetic and Biochemistry/ Faculty of Computing, Federal University of Uberlândia, Campus Patos de Minas, MG, Brazil.

Cucumber (*Cucumis sativus*) is one of the major vegetable crops worldwide, and in Brazil with an annual production that exceeds 200 tons. Its probable center of origin is in the mountainous regions of India, and there are currently more than 100 varieties of cultivable cucumber worldwide. Recent studies have shown the importance of gene regulation involving small RNAs, the processing system and their performance at the cellular level. MicroRNAs (miRNAs) are considered one of the most important small non-coding RNAs silencing mRNAs and controlling gene expression. There are multiple proteins responsible for the generation of miRNAs, among them the Dicer-like (DCL) protein stands out as one of the main protein. This work aimed to identify, using in silico analysis, genes involved in gene silencing pathway mediated by miRNAs in the genome of *C. sativus*. The sequences of the genome and transcriptome of *C. sativus* were obtained from public database Phytozome (<https://phytozome.jgi.doe.gov/pz/portal.htm>). The orthologous proteins from *Arabidopsis thaliana*, obtained from the RefSeq database of NCBI (<http://www.ncbi.nlm.nih.gov/>) were used as reference for the search using the BLASTp tool. The conserved domains and active site amino acid residues were recovered through the predicted proteins using the database PFAM and CDD. The phylogenetic analysis was conducted by Mega5.2 program. We were able to predict 17 DCL proteins divided into 4 subfamilies DCL1, DCL2, DCL3 and DCL4 in the genome of *C. sativus*. The putative protein DCLs were highly conserved at the amino acid level showing conserved domains such as RIBOc, PAZ super family, Dicer_dimer, DEXDc, Helicase_C, DSRM, Rnc and SrmB, these containing significant residues and conserved in critical positions on the protein. The putative active site of RIBOc domain was highly conserved at amino acid level showing the conserved amino acids glutamine, glutamic acid, aspartic acid / aspartic acid, glutamic acid in the similar positions when compared with orthologs proteins from plant species. The putative *C. sativus* DCL proteins were grouped in distinct clades in the phylogenetic tree corroborating the planned paralogs subfamilies. Thus, the results allow us to expand the study of miRNAs in cucumber, providing new challenges for understanding the biology of this organism.

Funding support: FAPEMIG

Archaeal RNA polymerase pausing modeling and its gene expression control impacts

Almeida-e-Silva DC , Tie Koide , Vencio RZN

Department of Computing and Mathematics FFCLRP-USP, University of São Paulo, Ribeirão Preto, Brazil

Department of Biochemistry and Immunology FMRP-USP, University of São Paulo, Ribeirão Preto, Brazil

Since the recognition of Archaea as a separate domain of life, there is growing interest in this evolutionary lineage. *Halobacterium salinarum* is an organism used as a model for halophilic archaea study. The NRC1 strain genome was published in 2000, and in 2008 the R1 strain sequencing, comparative analysis and proteomic data were published. These data were essential to the current understanding of haloarchaea as well as on archaeas as a whole.

Recently the massively parallel sequencing technology has enabled the identification of a new class of non-coding RNA, called the "Transcription Start Site-associated RNAs" (TSSaRNAs). The TSSaRNAs are small RNAs generated from the flanking regions of the Transcription Start Sites (TSS). TSSaRNAs have been identified in various eukaryotes and bacteria; and recent analyzes of *H. salinarum* data provide evidence of the presence of TSSaRNAs in Archaea. This suggests that TSSaRNAs are an evolutionarily widespread phenomenon in all three domains of life, and this phenomenon was possibly present in the last universal common ancestor. The conservation and expression regulation of non-coding RNA over all life domains suggests conserved biological functions associated with TSSaRNAs.

Currently there are not many confirmed information about the operation of TSSaRNAs. It is speculated that TSSaRNAs are related to the RNA polymerase pause sites; to elongation factors; to gene expression levels; or the backtracking process. In the present study we propose an *in silico* analysis of halophilic archaea data in search of answers to some of the unresolved issues that permeate the TSSaRNAs. With this we hope to advance the knowledge of this newly discovered class of small RNAs, in addition to expanding the understanding of archaea biology.

This work main goal is to advance on TSSaRNAs knowledge in archaea through *in silico* methods. For this, we intend to investigate sequence signatures that may be associated with RNA polymerase pause phenomenon; understand the role of the sequence signatures in archaea transcription; mathematically model the RNA polymerase pause phenomenon in archaea; establish relationships with observed phenomena and perform simulations seeking the relationship between RNA polymerase pause phenomena and gene expression control; investigate the influence of TSSaRNAs on transcription machinery fidelity and its possible consequences on the organism physiology.

In order to develop an automatic TSSaRNAs detection approach for *H. salinarum* and other organisms we are currently performing signal processing of RNA-seq, GC content, sequence conservation and minimal folding energy data.

Funding: Almeida-e-Silva is supported by a CAPES fellowship.

Transcriptome meta-analysis reveals the human organs evolution

Katia de Paiva Lopes, Ricardo Assunção Vialle, José Miguel Ortega

Departamento de Bioquímica e Imunologia, Instituto de Ciências Biológicas (ICB), Universidade Federal de Minas Gerais (UFMG), Brasil.

RNA-Seq allows the measurement of transcripts expression levels in a manner far more precise and global than previous methods. Studies using this technology have already altered our view of the extend and complexity of eukaryotic transcriptomes. Actually, multiple efforts have been made to determine and analyze the gene expression patterns of many human cell types in different conditions. However, until recently, little has been reported about the perspective of gene evolution and gene expression. So, in these work a transcriptome meta-analysis was performed using 4 different databases. We examined the protein coding genes with at least 1 FPKM by tissue from Fantom5, GTEx, HPA (Human Protein Atlas) and IBM (Illumina Body Map). This corresponds to the following number of genes, respectively: Brain (10164, 12137, 12501, 12049), Colon (9412, 12205, 12175, 11063), Heart (8424, 10333, 10794, 10070), Kidney (8584, 11961, 12084, 12002), Lung (9023, 12474, 12610, 11873), Ovary (8587, 11687, 11597, 12272), Prostate (8779, 12577, 12544, 12233) and Testis (9978, 14307, 14033, 13832). In a global analysis of each database, we classified the genes in some different groups, such as: Tissue Specific (TS - Expressed in only one tissue), Tissue Enriched (TE - mRNAs levels in a particular tissue at least five times those in all other tissues), Elevated genes (EG - Total number of tissues from another 3 groups as tissue enriched, group enriched and tissue-enhanced) and Ubiquitous (UB - Expressed in all tissues). The number of genes in these categories are: Fantom5 (1448, 1847, 10452, 4083), GTEx (1177, 2163, 11227, 6188), HPA (1504, 2762, 11208, 5922) and IBM (1966, 2932, 9640, 6617). Accessing the group of orthologues and determining the Last Common Ancestor of each gene, we determined the fraction shared among the clades of the human lineage, from Cellular organisms to Homo sapiens. Our results showed that the origin of distinct organs was genetically told by examining the time of appearance of some specific genes or within distinct groups as elevated genes. Analyzing the cumulative curve of genes originated in each clade showed that the area underneath the curve of TS genes is much lower than the area of UB genes for all databases analyzed here: Fantom5 (18.85, 24.91), GTEx (16.74, 25.06), HPA (16.95, 25.09) and IBM (17.67, 24.94), indicating that the ubiquitous genes are more ancient than specific genes. Moreover, when we used the prominent genes (most expressed genes) we verified a large intersection of it with ubiquitous or housekeeping genes, showing the same list of genes for different tissues.

Supported by: Capes, FAPEMIG, CNPq.

Transcriptome analysis of mice hearts infected with two strains of Trypanosoma cruzi: Insights into the parasite effects on the host gene expression

Castro TBR[§], Canesso MCC[§], Boroni ML*, Toledo NE#, Machado CR[§], Chiari E&, Macedo AM[§] and Franco GR[§]

*Departamento de Bioquímica e Imunologia[§], UFMG, Belo Horizonte/MG, Brazil;
Laboratório de Bioinformática e Biologia Computacional, Centro de Pesquisas, Instituto Nacional do Câncer*, Rio de Janeiro, Brazil; Departamento de Parasitologia, UFMG, Belo Horizonte/MG, Brazil&*

Chagas disease is a parasitic infection, caused by the kinetoplastid protozoan *Trypanosoma cruzi* (*T. cruzi*). Both, mammal host and *T. cruzi* parasites are highly genetically polymorphic, and therefore many variables come into play, making the disease outcome difficult to predict. For example, even more than a century after Chagas disease discovery, the mechanisms underlying the tissue tropism of *T. cruzi* has still to be elucidated. These complex interactions between host and pathogens comprise many still poorly understood surface and soluble glycoproteins as well as intracellular mediators. They are responsible to permit the survival of *T. cruzi* in the blood stream, adhesion to the host cellular membrane, as well as evasion to reactive oxygen species and immune system. On the other side, distinct MHC haplotypes in the mammal host may also change the disease course. In the present work, we seek to better understand the global host response or gene expression, to different *T. cruzi* strain parasites. To do so, we performed a RNA-Seq analysis of BALB/c mice hearts single or double-infected with two strains until day 15°, representing the acute phase of infection. Here, we demonstrate a clear distinction between the host gene expression against these two *T. cruzi* strains. Col1.7G2 (*T. cruzi* I), a known virulent strain, strongly activates Th1-polarized immune response genes. Whereas JG (*T. cruzi* II), a known non-virulent strain showed weaker expression of immunological genes while strongly inhibiting ribosomal proteins and oxidative pathways. Interestingly, enrichment pathway analysis of mice hearts infected with the mixture of the two strains, showed both phenotypes simultaneously. Altogether, our data aid to better understand the complex host-pathogen interactions in the context of Chagas disease, and the effect of different *T. cruzi* strains over mouse gene expression.

Comparative analysis of gene expression data between colorectal cancer cell lines with wild-type and silenced MMR genes

Cristóvão Antunes de Lanna, Nicole Scherer, Mariana Boroni

Laboratório de Bioinformática e Biologia Computacional (LBBC), Instituto Nacional de Câncer (INCA)

Colorectal cancer (CRC) is a high incidence carcinoma in Brazil and in the world, with high mortality rates in less developed countries. 34,280 new cases are estimated for 2016 in Brazil. Since the obtention of CRC samples at different stages of development is relatively easy, it allowed the in-depth study of CRC progression, as well as its molecular basis and the characterization of CRC subtypes. One of the known pathways of CRC progression is the microsatellite instability pathway (MSI), caused by mutations or epigenetic silencing in genes involved in mismatch repair mechanisms (MMR). As cell lines can be used to represent different stages of tumors according to their origin, this study aims to characterize the expression profiles of CRC cell lines with altered MMR pathway (MMR-) compared with control ones (MMR+). For this, RNA-seq data was obtained for nine CRC cell lines from project SRP052201 at the SRA database: MMR+ : SW480, HT29, COLO205, Caco2 and MMR- : LS174T, LoVo, HCT116, HCT15, RKO. After analyzing the quality of the reads and removing low-quality bases using FastQC and Trimmomatic, respectively, these sequences were aligned to the human genome (GRCh37 version) and sorted by coordinate, using the STAR algorithm. Then, the quality of the alignments was analyzed using RSeQC. After that, the mapped reads for each gene were quantified using HTSeq-count and the expression levels of genes were estimated through the edgeR tool, written in R, from the Bioconductor repository. A total of 245 differentially expressed genes were obtained, of which 229 were down-regulated and 16 were up-regulated in the MMR- cell lines. Differentially expressed genes were classified according to the Gene Ontology (GO) database categories, using the clusterProfile Bioconductor package. Enriched biological function categories include processes related to angiogenesis, response to wounding, and lipid and steroids metabolism. Membrane and extracellular matrix components represented the majority of cellular components enriched in this analysis. Enriched categories for molecular function included sulfur compound, heparin, and glycosaminoglycan binding.

Functional and structural characterization of RBP42 in *Trypanosoma cruzi*

Daniela de Laet Souza¹, Daniela Ferreira Chame¹, Eddie Luidy Imada¹,
Helaine Graziele Santos Vieira², Andrea Mara Macedo¹, Carlos Renato Machado¹,
Dominik Kaczorowsk², Glória Regina Franco¹

1- Departamento de Bioquímica e Imunologia, UFMG, Belo Horizonte/Brazil,

2- Garvan Institute of Medical Research , Sydney/Australia

Trypanosoma cruzi, the etiological agent of Chagas disease, has unique characteristics in genome architecture and gene expression regulation. In this parasite, genes of unrelated functions are transcribed as long polycistronic pre-mRNAs and solved into monocistronic transcripts by the trans-splicing and polyadenylation processes. Due to these unique characteristics, gene expression regulation occurs mainly at the post-transcriptional level by RNA binding proteins (RBPs) that orchestrate transcripts processing, transportation, stabilization and degradation under normal and stress conditions. Moreover, there is a large variety of RBPs coding genes in the parasite's genome. The interaction between RBPs and mRNAs forms structures called ribonucleoprotein complexes (RNPs) that can aggregate into microscopically visible cytoplasmic structures known as RNA granules in response to stress. RNA granules could function as centers for degradation and storage of transcripts used in stress recover. To investigate the role of RBPs, epimastigotes of the CL Brener clone were transfected with the vector pRock_Neo to generate a cell line overexpressing TcRBP42. The phenotype of the transfected cells was characterized through growth curves in cells treated or non-treated with UV light, benznidazole and gamma radiation. The overexpression of the TcRBP42 in transfected cells was confirmed by detection of the histidine-tagged RBP42 in whole protein extracts by Western blot. The analyze of growth curves revealed a similar growth pattern between RBP42 overexpressing cells and control cells at normal conditions. In contrast, these cells were more sensitive to UV light (1000 J/m² dose) and more resistant to gamma radiation. It was also observed that the RBP42 overexpressing cells are also more sensitive to benznidazole treatment. Given that RBP42 gene is annotated in the *T. cruzi* genome as a hypothetical protein, functional and structural analyses were performed to predict conserved domains and secondary structure by using the softwares InterProScan and PSSpred. In silico analyses revealed the presence of NTF-2 and RRM domains and a secondary structure similar to the one characterized for *T. brucei*. The RBP42 levels of gene expression were analyzed by RNAseq (package DESeq2 from R/Bioconductor) in non-transfected cells 4, 24 and 96 hours after exposure to 500 Gy of gamma radiation. The RBP42 gene is more expressed than the other genes in its genome vicinity. It was observed a two-fold change increase in the RBP42 gene expression levels after gamma irradiation, suggesting a role in the parasite stress response.

RNA-binding proteins ALBA3 and DRBD3 characterization on *Trypanosoma cruzi* under gamma irradiation stress

Daniela Ferreira Chame¹, Daniela De Laet Souza¹, Eddie Luidy Imada¹, Helaine Graziele Santos Vieira², Andrea Mara Macedo¹, Carlos Renato Machado¹, Dominik Kaczorowski², Glória Regina Franco¹

¹Departamento de Bioquímica e Imunologia, UFMG BH/Brazil, ²Garvan Institute of Medical Research, Sydney/Australia

Trypanosoma cruzi is highly resistant to gamma irradiation. After 500 Gy of ionizing radiation dose the chromosomal bands are fragmented, but after 48 hours the parasite is able to restore its initial chromosomal patterns. RNA binding proteins (RBPs) are important modulators of gene expression in normal and stress conditions and are involved in processing, decay, stability and transport of mRNAs. In order to investigate the involvement of RBPs in the response to radiation stress, we aimed to characterize the *T. cruzi* RBPs TcALBA and TcDRBD3 and their associated RNAs in irradiated and non-irradiated parasites. To explore the role of these RBPs, epimastigotes of the CL Brener clone were transfected with the vector pRock/Neo to obtain *T. cruzi* cell lines overexpressing the proteins TcALBA and TcDRBD3, containing a C-terminal 6His tail. The recombinant proteins were successfully detected by western blot in whole protein extracts from transfected epimastigotes. The phenotype of RBPs overexpressing cell lines was evaluated through growth curves in the presence and absence of stress induction. Under gamma radiation stress, epimastigotes overexpressing TcALBA showed a faster growth recovery while TcDRBD3 showed a slower growth recovery in comparison with control cells. Immunofluorescence assays were performed to visualize the distribution pattern of these proteins throughout the cell. TcALBA3 was mainly located in the cytoplasm, 4h and 24h after irradiation. On the other hand, TcDRBD3 exhibited a perinuclear localization 4h after irradiation and a cytoplasmic granular distribution, 24h after irradiation. Given that TcALBA3 and TcDRBD3 genes are annotated in the *T. cruzi* genome as hypothetical proteins, structural analyses of these proteins were performed to predict conserved domains and secondary structures using the softwares InterProScan and PSSpred. *In silico* analyses revealed the presence of ALBA and RRM domains on TcALBA3 and TcDRBD3 and a secondary structure similar to *T. brucei* ALBA3 and DRBD3, respectively. TcALBA3 and TcDRBD3 gene expression levels were also analyzed by RNAseq (package DESeq2 from R/Bioconductor) in non-transfected cells on 4, 24 and 96 hours after exposure to 500 Gy of gamma radiation. TcALBA3 and TcDRBD3 are among the most expressed genes in their genomic vicinity. It was observed a two-fold change reduction in TcDRBD3 and TcALBA after gamma irradiation, suggesting their role in the parasite stress response. In order to continue our investigation on the function of RBPs in response to gamma radiation, we intend to perform RIP assays and identify RNAs present in these RBPs complexes.

Occurrence of differential alternative splicing in the transcriptome of mice hearts infected with two populations of *Trypanosoma cruzi*

Toledo NE¹, Castro TB¹, Machado CR¹, Rodrigues NA¹, Viana A², Chiari E²,
Macedo AM¹, Franco GR¹

Department of Biochemistry and Immunology-ICB/UFMG , Department of Parasitology-ICB/UFMG

Since the description of Chagas disease, caused by the protozoan parasite *Trypanosoma cruzi*, the mechanism underlying the parasite tissue tropism has yet to be revealed. Our group has previously shown that different strains of *T. cruzi* (JG and Col1.7G2) had a differential tissue tropism in BALB/c mice upon infection., being JG preferentially found in hearts and Col1.7G2 in other tissues of these animals, especially in the rectum. Transcriptome sequencing (RNA-seq) of mRNA extracted from BALB/c infected hearts (groups: JG, Col1.7G2 and an equivalent mixture of both strains) was compared to non-infected mice and showed a predominance of upregulated genes in Col1.7G2-infected animals. In the other side, JG-infected mice had a great number of downregulated genes. Curiously, the mixture-infected group showed both cases simultaneously. Alternative splicing is a RNA processing in which different exons and introns of the same pre-mRNA may be skipped or retained to produce different mature mRNAs, largely expanding the transcriptome repertoire. Thus, the aim of this study was to evaluate the occurrence of differential alternative splicing in the transcriptome of mice infected with both *T. cruzi* strains. For initial analyses of the transcriptomes, the quality of sequences was accessed with the FastQC software. Subsequently, we performed alignment against the mouse reference genome using the splice-aware aligner, STAR. After alignment, the program Multivariate Analysis of Transcripts Splicing (rMATS) was used for recognition of the main types of alternative splicing patterns namely exon skipping (ES), mutually exclusive exons (MXE), alternative 5' splice site (A5SS), alternative 3' splice site (A3SS) and intron retention (RI). Our present result showed that intron retention and exon skipping have less inclusion level in Col1.7G2 and mixture-infected group mice when compared to the control group and only alternative 3' splice site prevailed in mice hearts with mixed infection. Comparing JG infected mice with the control group only exon skipping have higher inclusion level in the infected group over the others events. In conclusion, we have shown that, in the experimental model of Chagas disease, different *T. cruzi* strains can remodel the splicing pattern of the host and this may be relevant for disease development and the parasite tissue tropism.

Long Noncoding RNAs in Patients with Dengue: Insights into Gene Regulation

Matheus C. Bürger^{1,2}, Lucas E. Cardozo¹, Thiago D. C. Hirata¹, Helder T. I.

Nakaya^{1,2}

1 - School of Pharmaceutical Sciences – University of São Paulo, São Paulo, SP, Brasil; 2 - Graduate Program in Bioinformatics - Institute of Mathematics and Statistics - University of São Paulo, SP, Brasil;

Clinical manifestations of dengue viral infections may vary from fever to the potentially deadly dengue hemorrhagic fever and dengue shock syndrome. Several studies have been published investigating global gene expression changes between healthy subjects and dengue-infected patients with different clinical manifestations. However, none of these studies have analyzed the putative role of long noncoding RNAs (lncRNAs) during these conditions. Here, we performed a meta-analysis using three publicly available microarray datasets of dengue-infected patients, focusing on lncRNA expression and their potential mechanisms in gene regulation. In order to identify probes that represent lncRNAs, we have reannotated all major commercial microarray platforms. The reannotation consisted in cross-referencing the genomic coordinates of microarray probe sequences to the gene sequences of the following databases: Gencode, Noncode, LNCipedia and MiTranscriptome. By doing this, we identified 6603 probes that represent potential lncRNAs. We then re-analyzed three studies from the Gene Expression Omnibus database (GSE18090, GSE43777 and GSE51808) composed of 60 control samples, 66 uncomplicated dengue patients and 60 severe dengue patients in total. Differential expression analyses using LIMMA package revealed that the expression of hundreds of transcripts were consistently altered between patients infected with dengue and healthy subjects in the three studies (p -value < 0.005 and absolute log₂ of fold-change > 0.32). Of those, 12 were annotated as lncRNAs. One long noncoding RNA with lower expression in Dengue is TUG1, which was previously described as an important gene in different types of cancer. TUG1 can bind to Polycomb Repressive Complex 2, which is known to lead to chromatin remodulation. Evidence of cis-regulation was found through the calculation of Spearman correlation coefficient between the expression of antisense lncRNAs and their sense protein-coding transcripts. This "guilty-by-association" approach suggests that, during Dengue infection, the USP30-AS1 lncRNA can modulate the alternative splicing of USP30 gene, which in turn is important for mitochondrial deubiquitination and there are several studies showing the importance of ubiquitination system in viral infections. Taken together, our analyses revealed that lncRNAs can play an important role in the immune responses to dengue infection. Financial Support: CAPES.

Transcriptome analysis of *Corynebacterium pseudotuberculosis* in an iron deficient environment

Ibraim, I.C.¹; Aburjaile, F.F.¹; Castro, T.L.P.¹; Pinto, A.C.¹; Seyffert, N.¹; Souza, E.M²; Azevedo, V¹.

¹Laboratório de Genética Celular e Molecular, Instituto de Ciências Biológicas, UFMG, Belo Horizonte, Brasil. ²Universidade Federal do Paraná, Setor de Ciências Biológicas, Departamento de Bioquímica, Paraná, Brasil
Email: Izabelaibraim@msn.com

Livestock is one of the fastest growing sectors of the agricultural economy and is boosted by high profitability and technological advances. The sector growth offers development opportunities, poverty reduction and nutritional gains. Nevertheless, one of the limiting factors of this activity is the high prevalence of infectious diseases that affects the flocks, thus reducing viability and exportation, raising costs and reducing profitability. In this context, infections by *Corynebacterium pseudotuberculosis* are amongst the most relevant and their occurrence is related to high economic impact diseases. "Omics" studies regarding this pathogen have enabled the identification of putative pathogenic islands containing classical virulence elements, including genes involved in iron uptake. For many pathogenic bacteria, iron availability and uptake contributes to a successful host colonization and bacterial survival. Due to its importance and the fact that mammalian host species restrict the iron availability to control bacterial infection, the interference with iron-acquisition mechanism can be used as a target for the development of novel antimicrobial treatments as well as a more effective vaccine therapy. Still, although its critical relevance, the mechanism involved in the uptake, virulence and availability of iron in the *C. pseudotuberculosis* species is poorly understood. Furthermore, large scale RNA sequencing (RNA-seq) technology can be used to expand our comprehension of the functional genomics involved in the infection, resistance and survival of this important pathogen. In this context, the aim of this project is to characterize the RNA response of two strains of *Corynebacterium*: CpT1 (wildtype) and Cp13 (deficient iron transport binding protein - *ciuA* mutant), in a low iron environment. In order to achieve a low iron environment, different concentrations of Fe²⁺ chelator, 2,2'-bipyridine (bipyridyl, BIP), were used to analyze bacterial growth in BHI medium. Low, nonfatal iron concentration, was achieved with 250uM of BIP, RNA extraction and purification was done 6 hours after inoculation. RNA-seq is going to be carried out using the Ion Proton platform and analyses will include quality assessment through FASTQC, filtering of high quality transcripts and alignment to a reference genome with the software TopHat v2.1.0. Single aligned reads will be quantified using HTSeq counter and the differential gene expression statistical analysis will be done using the EdgeR/Bioconductor package. Although still under development, the hypothesis is that low environmental iron concentration could trigger a switch in the expression of *Corynebacterium pseudotuberculosis* genes involved in host colonization and persistence, including those associated with iron acquisition and virulence factors.

Financial Support: CNPq and Capes.

Annotation and analysis of the dynamics of splice acceptor sites in *Trypanosoma cruzi* under gamma radiation stress

Reis, A. L. M.¹; Bitar, M.²; Vieira, H. G. S.³; Kaczorowski, D.³, Macedo, A. M.¹; Machado, C. R.¹; Franco, G. R.¹;

1. Departamento de Bioquímica e Imunologia, UFMG, BH/MG, Brazil;

2. QIMR Berghofer, Brisbane, Australia; 3. Garvan Institute of Medical Research, Sydney, Australia

Unlike most eukaryotes, in Trypanosomatids, genes are not interrupted by introns and transcription is polycistronic. Maturation of individual mRNAs is accomplished by coupling spliced leader trans-splicing (SLTS) and polyadenylation. The attachment of a 39-nucleotide sequence, the spliced leader, to the 5' end of individual cistrons upon recognition of splice acceptor sites resolves the 5'UTR. Alternative SLTS has not been characterized in *Trypanosoma cruzi*, although different regulatory roles have been speculated for this mechanism. A deeper investigation is necessary to clarify how specific splice acceptor sites are selected under diverse environmental conditions. This parasite is highly resistant to different sources of stress, sustaining an exposure to 500 Gy of gamma radiation. Under this circumstance, genomic DNA is fragmented, but the karyotype is gradually restored, leading to a complete re-establishment of the chromosomal band pattern in less than 48 hours. The aim of this study is to characterize trans-splicing at structural and functional levels, observing its dynamics after *T. cruzi* is exposed to ionizing radiation stress. A time-series experiment was designed to evaluate changes in the transcriptome of epimastigotes of the CL Brener strain not exposed versus 4, 24 and 96 hours after exposure to 500 Gy of gamma radiation with two biological replicates at each time point. Following total RNA extraction, libraries were prepared with Truseq mRNA Stranded and then sequenced in the Illumina Hiseq2500 platform. The pipeline applied for the annotation of splice acceptor sites was mainly comprised of: FastQC (for quality check), Cutadapt (for identifying and trimming the spliced leader sequence), BWA-mem (for mapping to the reference genomes), Python in-house scripts (for the actual calling of splice sites) and R scripts (for statistical analyses). A total of 48,719 different splice sites were identified using this protocol. They were assigned to 17,053 annotated genes, while there were 5,658 genomic regions with splice sites but no genes annotated. Further inspection showed that 1,384 genes out of those annotated may have a wrong CDS start and 1,228 new CDSs were found in those regions with no annotation. Most of the sites are present in the control (~71%) and there is a clear distinction, in terms of coverage, between the conditions non-irradiated versus irradiated. Some genes even have a change in their main splice site over time. In summary, this work enables a refinement of genome annotation and a better understanding of the dynamics of SLTS in *T. cruzi*.

De novo assembly of Trypanosoma cruzi strain CL Brener transcriptome

Eddie Luidy Imada¹, Mainá Bitar², Maíra Ribeiro Rodrigues¹, Daniela Ferreira Chame¹, Helaine Grazielle dos Santos Vieira³, Michele Araújo Pereira¹, André Martins Reis¹, Dominik Kaczorowski³, Willian Santos Prado¹, Andréa M. Macedo¹, Carlos R. Machado¹, Martin A. Smith³, Glória R. Franco¹

¹Departamento de Bioquímica e Imunologia, UFMG, Belo Horizonte – Brazil, ²QIMR Berghofer Medical Research Institute, Brisbane – Australia, ³Garvan Institute of Medical Research, Sydney – Australia

Chagas disease is a neglected tropical disease caused by *Trypanosoma cruzi* that is estimated to affect at least 6 million people worldwide. The first genome draft of the hybrid clone CL Brener was published in 2005 as several scaffolds and contigs, which were latter partially assembled into 82 chromosomes in mid 2009. Despite the great improvement of the *T. cruzi* genome assembly it still presented many gaps and unplaced contigs. Another caveat of the current genome is that its current annotation are almost exclusively based on automatic ORF detection, which might skew transcriptome analysis due to the lack of unannotated genomic elements. To address these problems we have assembled a *de novo* transcriptome of *T. cruzi* CL Brener using ~70 million paired end Illumina Hiseq 2500 reads with Trinity. The assembly resulted in 57,084 transcripts that were clustered using CD-HIT-EST into 24,844 non-redundant clusters. The clusters representatives (herein referred simply as transcripts) were annotated using a 2-steps methodology by first searching for homologues with BLASTN at nucleotide level against current predicted CL Brener transcripts and then searching at protein level with BLASTX against the TriTrypDB for those failing to align at nucleotide level. As a result, 93% of the transcripts were assigned to 14883 annotations, with 6038 genes being completely covered and 8,665 genes assembled by at least half of its predict length. Transcripts expression evaluation with Kallisto and GO annotation showed that the most expressed genes were related to basic metabolism as expected since steady epimastigote cultures were used in this work. Furthermore, using the assembled transcriptome we were able to close over 400 gaps in the current genome, improve/create UTR annotations for 9206 genes and correct some misplaced contigs in the assembled chromosomes. Our approach shows how *de novo* transcriptomes can be leveraged not only to functional studies, but to improve genome assemblies and annotations as well.

Financial Support: CAPES, FAPEMIG

Evaluating RNA single bulges with a mesoscopic model

Erik de Oliveira Martins, Gerald Weber

Department of Physics, Federal University of Minas Gerais, Belo Horizonte, MG, Brazil

Thermodynamic models at the mesoscopic scale such as the Peyrard-Bishop model are capable of reinterpreting experimental measurements in terms of intramolecular interactions. For instance we recently were able to obtain detailed information about hydrogen bonds for the GU wobble pair in RNA. For example the distinctive single hydrogen bond that GU may assume under certain circumstances was correctly predicted with a mesoscopic models. Here we address the far more complex situation of single bulges in RNA. In this case a single nucleotide is left unpaired which may then either flip out or interact with neighboring intrastrand bases through $\pi - \pi$ interaction. A bulge may in some cases also disrupt neighboring base pairs. Representing bulges with mesoscopic approaches represents a challenge as such models were developed for unperturbed helical structures. We adapted successfully the Peyrard-Bishop Hamiltonian to represent the perturbation induced by a single bulge, in particular we were able to treat the resulting asymmetry of the stacking interaction. We are now determining the optimal parameters for this model, for this we used from the literature 133 melting temperature sequence data of A, C, G and U bulges surrounded by AU and CG base pairs. Our first results indicate that for some sequence contexts the hydrogen bonds of neighboring base pairs tends to vanishing values indicating that a base pair rupture has occurred. However, for most contexts the flanking base pairs retain a very stable hydrogen bond. We are also addressing the problem of location of type II bulges. In this case the exact pairing of the bulge is not known. Our parametrization procedure will cover the various location possibilities and work out the most stable location for the type II bulge to reside in the sequence. We acknowledge financial support by Capes, CNPq and Fapemig.

Study on the variation of RNA Secondary Structure prediction as a function of Thermodynamic Parameters

Rodolfo Vieira Maximiano¹², Gerald Weber²

¹*Centro Federal de Educação Tecnológica de Minas Gerais - unidade Contagem,*

²*Departamento de Física - Universidade Federal de Minas Gerais*

In this work we evaluated how the accuracy of RNA secondary structure prediction softwares is modified as a function of the thermodynamic parameters. The parameters in use today are obtained from DNA and RNA melting experiments and each measurement is subject to an intrinsic experimental error. These experiments are modeled using the nearest-neighbour technique, which is one of the most commonly used theoretical models in this field of study. Here we used the RNAFold software from the Vienna package, which is in frequent use for bioinformatics applications. Our purpose was to verify the impact of the measured experimental error in the software ability to predict the final RNA secondary structure of a diverse database of known RNAs, which have folded into structures determined experimentally by other methods. When comparing both the experimentally known and predicted structures we can evaluate the prediction quality. The indicators of prediction quality used in this work are often used in literature, namely, the positive predict value (PPV) and the sensitivity. We created additional sets of nearest-neighbor thermodynamic parameters taking into account the reported experimental error. For each new set, we performed the folding of all known RNA sequences, calculating a general PPV and sensitivity quality indicator for that particular set. We are also performing verifications of the variation of prediction accuracy when subdividing the known RNA sequences by size, with the intent of detecting how, and when, does the prediction ability of the software cease to be as precise as it is for small sequences. So far our results indicate a stronger modification in the prediction accuracy when modifying the hairpin and bulge formation parameters. In comparison, modifying the base pair stacking parameters has generally a lesser impact on the folding prediction quality.

Funding Agencies: CNPQ, FAPEMIG, CAPES

Analysis and comparison of force field of RNA using molecular dynamic simulations.

Rodrigo Bentes Kato , Jadson Cláudio Belchior

Departamento de Bioinformática da Universidade Federal de Minas Gerais, Departamento de Química da Universidade Federal de Minas Gerais

DNA and RNA are big and flexible polymers that be choice by nature to transmit information. The most common 3D structure is represented by the helix, but these biopolymers are extremely flexible and polymorphic. They can easily change its structure to adapt to different interactions and purposes. RNA has numerous key cellular roles in addition to being the intermediate molecule for the gene expression, as the genetic information stored in DNA is decoded to proteins. In the literature, the proteins remain the most characterized cellular effectors, whilst the pivotal roles performed by RNAs within the cell are more recently being fully appreciated. Despite the great importance of nucleic acid–protein interactions in the cell, our understanding of their physico-chemical basis remains incomplete. Our understanding of the importance of these unusual or transient structures is growing, as recent studies of RNA topology, supercoiling, knotting and linking have shown that the geometric changes can drive, or strongly influence, the interactions between protein and RNA, so altering its own metabolism. In order to address this challenge, we used molecular dynamics simulations (DM) to analyze some standard force fields and new refinement of force field applied in nucleobases of RNA. A comparison against set of available values in the literature and experimental data attests to the quality of the computational approach and the force field. This work is a crucial help in the understanding and planning of natural and artificial nanostructures is given by modern computer simulation techniques, which are able to provide a reliable structural and dynamic description of nucleic acids. The force field will improve usage in various practical applications such as docking, interface design and structure prediction.

Transcriptome profiling in *Leishmania amazonensis* promastigotes associated with virulence attenuation

Gabriela Flavia Rodrigues-Luiz¹, Mariana Costa Duarte², Daniel Menezes-Souza², Ricardo Toshio Fujiwara¹, Eduardo Antonio Ferraz Coelho², Daniella Castanheira Bartholomeu¹

¹*Instituto de Ciencias Biologicas Universidade Federal de Minas Gerais, ²Colegio Tecnico Universidade Federal de Minas Gerais*

Leishmaniasis is one of the most important neglected tropical diseases and it is known that in vitro cultivation of *Leishmania* spp. for long periods results in a progressive loss of virulence. The focus of this work was to integrate -omic data with bioinformatics resources to contribute to a better understanding of an important biological aspect of this parasite: the loss of virulence after successive periods of *in vitro* cultivation. For this purpose, we evaluated by RNA-seq the difference in expression profile of *L. amazonensis* promastigotes freshly isolated from experimentally infected mice (R0) and parasites that were cultured after 30 passages *in vitro* in Schneider's Insect Medium (R30). We have identified 683 genes with significant differential expression, 64.12% of which with decreased expression in R30 compared with R0. This study showed that the loss of virulence in *L. amazonensis* after successive periods of *in vitro* cultivation are likely to be associated with parasite-host interactions mediated by parasite surface proteins, stress tolerance and metabolism of amino acids and fatty acids. Furthermore, we disclosed several other genes that are possibly associated with Leishmania virulence and are good candidates for further functional studies. In this study we have also investigate the presence of viral sequences in the *L. amazonensis* RNA-seq reads. To this end, we assembled reads that were not mapped against the *Leishmania amazonensis* reference genome using the Trinity software and performed a Blast search against the NCBI non-redundant database. The results were manually filtered by length and e-value. We identified 35 putative viral unigenes and, based on their sequence similarity, the sequences belong to Picornavirales order and Baculoviridae family. Further studies are necessary to confirm the identity and phylogeny of these putative virus sequences. Once the virus identity is confirmed, the impact of these viruses on the virulence of *L. amazonensis* will be investigated.

De novo transcriptome assembly and comparative expression profiling of midgut tissues of four non-model insects

Rajesh Kumar Gazara¹, Christiane Cardoso², Daniel Bellieny-Rabelo¹, Clélia Ferreira², Walter R. Terra², and Thiago Motta Venancio¹

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, Brazil.

2 Departamento de Bioquímica, Instituto de Química, Universidade de São Paulo, São Paulo, Brazil.

In the present work we describe the *de novo* transcriptome assembly and expression profiling of midgut tissues from four non-model insect species belonging to different orders: Lepidoptera (sp. *Spodoptera frugiperda*, a maize pest insect), Diptera (sp. *Musca domestica*, house fly, transmits human and animal diseases), Hemiptera (sp. *Dysdercus peruvianus*, a major cotton pest) and Coleoptera (sp. *Tenebrio molitor*, a storage grain pest). Total mRNA samples of posterior midgut, anterior midgut, whole midgut and carcass (i.e. body without midgut tissues) samples were submitted to pyrosequencing using a 454 instrument. Sequencing reads were filtered to remove low -quality and contaminant reads and assembled *de novo* with different algorithms. Different assemblies obtained with MIRA and Newbler were merged with CAP3, as this strategy has been previously shown to provide better results than the usage of a single algorithm. We obtained 6,395, 9,010, 4,005 and 6,833 unigenes for *D. peruvianus*, *T. molitor*, *M. domestica* and *S. frugiperda*. Differentially expressed genes were inferred by comparing different tissues within each species. A stepwise strategy was employed to functionally annotate unigenes. In total, 79.69 to 93.09% of unigenes were assigned with some functional information. A total of 2,970, 5,964, 2,506 and 3,627 unigenes from *D. peruvianus*, *T. molitor*, *M. domestica* and *S. frugiperda* were assigned to eggNOG orthologous groups (OGs), respectively. OG information was used to find functions that are commonly enriched among highly expressed genes the same tissues across different species, as well as to find gene families with divergent expression patterns across species and tissues. Particular gene families were explored in more detail, such as important digestive enzymes. Taken together, our results provide a collection of genes and gene families with critical roles in the development, maintenance and biochemistry of midgut tissues across divergent insect species.

Funding: The authors acknowledge Universidade Estadual do Norte Fluminense Darcy Ribeiro and the following Brazilian funding agencies for their support: A Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES).

Differential expression in colorectal cancer progression

Aline Duarte Gomes, Paulo Thiago Santos, Nicole Scherer, Mariana Boroni

Laboratório de Bioinformática e Biologia Computacional - Instituto Nacional de Câncer (INCA), Rio de Janeiro, RJ, Brazil

Colorectal cancer (CRC) is the third worldwide most commonly diagnosed cancer in men and the second one in women, with more than 1 million cases per year. Deaths from CRC are relatively lower (8.5% of the total) with more deaths (52%) in the less developed regions of the world, reflecting a poorer survival in these regions. Estimates for Brazil indicate that CRC will affect more than 34,000 people in 2016. Furthermore, CRC provides a good model for the study of morphological and genetic stages in cancer progression, since its tumor progression process is very well described. Large-scale studies have pointed out genomic, epigenomic and gene expression alterations that contribute to this tumor development. However, each of these studies provides an one-dimensional and limited view of this whole system. We aim to study the differential gene expression in CRC during its progression and investigate important pathways affected in different stages. In order to identify differentially expressed genes, we have analyzed RNA-seq data from the TCGA database. Our data set was composed by 287 colon cancer and 41 normal tissue samples (14 of which are paired), totaling 328 samples. For the differential expression analysis, we employed DESeq2 package from Bioconductor for R, with a cutoff values set to $P < 0.001$ and $\log_2\text{FoldChange} \geq 1.58$. Then, using the package clusterProfiler from Bioconductor repository, we identified the enriched pathways in the Gene Ontology (GO) database in differentially expressed genes with a cutoff set to 0.05 for both pvalue and qvalue. A total of 2,823 differentially expressed genes were found, of which 1,158 are up-regulated and 1,594 are down-regulated. We found that the differentially expressed genes are enriched in high activity channels, transmembrane transporters, receptors and growth factor in molecular function categorie; for biological process, we see an impact on the muscular system, on the cellular homeostasis processes, in regulating hormone levels and blood circulation; and the terms enriched in cellular component indicate alterations in the extracellular matrix, cellular transport, membrane components and components of the muscular system. These data shows important processes that are affected in CRC, as higher expression of receptors, growth factors and hormones that could be related to proliferative signaling, and also alterations in the extracellular matrix that could be important to the activation of invasion and metastasis, important hallmarks of cancer.

Keywords: colorectal cancer, gene expression, RNA-seq, TCGA

Comparative transcriptome profiling of virulent and non-virulent *Trypanosoma cruzi* underlines a role of surface proteins during infection

Gabriela F. Rodrigues-Luiz^{1*}, A. Trey Belew^{2*}, Rondon P. Mendonça-Neto³, Bruna M. Valente³, Antonio Edson R. Oliveira³, Rafael B. Polidoro⁴, Ricardo T. Gazzinelli^{3,4}, Daniella C. Bartholomeu¹, Barbara A. Burleigh⁵, Najib M. El-Sayed^{2#} and Santuza M.R. Teixeira^{3#}

¹Departamento de Parasitologia, Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brazil, ²Department of Cell Biology and Molecular Genetics, University of Maryland, College Park, MD, US, ³Departamento de Bioquímica e Imunologia, Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brazil, ⁴Fundaçao Oswaldo Cruz/Centro de Pesquisa Rene Rachou, Belo Horizonte, MG, Brazil and ⁵Department of Immunology and Infectious Diseases, Harvard School of Public Health, Boston, MS, US.

Trypanosoma cruzi, the protozoan that causes Chagas disease, has a complex life cycle involving several morphologically and biochemically distinct stages that establish intricate interactions with various insect and mammalian hosts. It has also a heterogeneous population comprising strains that shows distinct properties such as virulence, sensitivity to drugs, antigenic profile and tissue tropism. We analyzed transcriptome data from two cloned *T. cruzi* strains that display contrasting virulence phenotypes in animal models of infection: CL Brener is a virulent strain and CL-14, a strain that is neither infective nor pathogenic in in vivo models of infection. RNA-seq analysis of CL Brener epimastigotes, trypomastigotes and intracellular amastigotes harvested at 60 and 96 hours post-infection (hpi) of human fibroblasts revealed large differences in their gene expression profiles. These changes reflect the parasite's adaptation to distinct environments during the infection of the insect vector and mammalian cells, including changes in energy sources, oxidative stress responses, cell cycle control and cell surface components. Whereas an extensive transcriptome remodeling was observed when CL Brener trypomastigotes were compared to 60 hpi amastigotes, only minor differences were observed between 96 hpi amastigotes and trypomastigotes of CL Brener. In contrast, the differentiation of the avirulent CL-14 from 96 hpi amastigotes to trypomastigotes was associated with considerable differences in gene expression particularly in genes encoding surface proteins such as trans-sialidases and the mucin associated surface proteins (MASPs). Thus, our comparative transcriptome analysis indicates that the avirulent phenotype of CL-14 may be due, at least in part, to a reduced or delayed expression of genes encoding surface proteins that are associated with the transition of amastigotes to trypomastigotes, an essential step in the establishment of the infection in the mammalian host.

Acute Myeloid Leukemia gene co-expression networks and differential expression analysis in blood and bone marrow samples

Kendi Nishino Miyamoto, Diego Bonatto

Centro de Biotecnologia, Universidade Federal do Rio Grande do Sul

Acute myeloid leukemia (AML) is one of most aggressive hematological diseases, characterized by abnormal growth of immature myeloid hematopoietic cells inside the bone marrow (BM) and subsequent organ and tissue invasion via bloodstream. This causes a myriad of symptoms that goes from fever and shortness of breath to neurological manifestations and severe hemorrhage that could lead to death if there is no BM transplant or chemotherapy available or effective. Studies demonstrated that cancer cell environment is crucial to determine its growth and development. In this sense, the transcriptional profile of leukemic cells could allow to understand how these cancer cells are able to grow and disseminate quickly along the BM and the bloodstream. Thus, blood and BM RNA-seq data from AML patients were obtained from the Gene Expression Omnibus database, which were quality assessed and pre-processed using FastQC and Trimmomatic software. Trimmed reads were aligned to Ensembl reference transcriptome (version GRCh38.84) using Salmon software. Gene quantification were acquired based on transcripts count and used as an input to perform differential gene expression (DGE) and to build a co-expression network (CN) through DESeq2 and WGCNA R packages, respectively. Results showed a total of 1426 DEG (Blood x BM), which 416 and 1010 were over and underexpressed, respectively and they are mostly related to cell division, cell adhesion and also immune response and signaling processes mediated by immunoglobulins and cytokines. CNs showed, modules in both samples that have common biological processes, like immune response, cell cycle and DNA repair pathways. However, they differ in their composition and structure, as well as some of them are linked to different pathways. These preliminary results showed some possible insights about how these cells behave in different environments, although more analysis in these modules are required to understand, in a broad sense, how it can affect tumor biology.

Large scale transcriptional analysis of an animal model of seizures

Samara Damasceno, Cristiane de Souza Rocha, Iscia Teresinha Lopes-Cendes, Ana
Lúcia Brunialt Godard

Instituto de Ciências Biológicas-UFGM and Faculdade de Ciências Médicas-UNICAMP

Wistar Audiogenic Rat (WAR) is an animal model in which seizures are developed by acoustic stimulation that activate the quadrigeminal plate structure. The characterization of the model's transcriptional profile can elucidate aspects regarding gene regulation related with the occurrence of seizures. The purpose of the study was to evaluate the WAR quadrigeminal plate transcriptome in post-ictal state. Two Wistar and two WAR rats were submitted to acoustic stimulation. The WAR exhibited seizures and the Wistar did not respond to the stimulus. Four days after the stimulus, the quadrigeminal plates were collected and processed to prepare the libraries. The RNA-Seq was realized in MiSeq platform (Illumina). The screened data by Cutadapt software were mapped using the Bowtie2 software. The gene counting was made at the HTseq and Features Counts and the differential expression was determined using DEseq and EdgeR softwares. The PCA was realized in R, the genes were submitted for functional annotation by GO Consortium and, finally, the results validation was performed by qPCR. Considering the value FDR $\leq 0,05$ by EdgeR, 62 genes were identified differentially expressed between the WAR and the control. DEseq identified 16 genes, in which 14 genes were also recognized by EdgeR. Considering both analysis, 28 genes were upregulated and 36 downregulated in WAR. The PCA revealed a segregation between the samples of Wistar and WAR, showing that the seizure's predisposition is the main determinant of gene expression variation between these groups. The functional annotation clustered the genes in six categories of molecular functions: binding, catalytic activity, receptor, signal transducer, structural molecule and transporter. Among the 16 genes identified by DEseq, 13 were validated by qPCR. Three genes, *Gpr126*, *Gria2* (receptor category) and *Qdpr* (catalytic category) showed an interesting result regarding expression pattern and the connection of their function with the phenotype. The *Gpr126* gene was downregulated and the *Gria2* and *Qdpr* genes were upregulated in WAR animals that had seizures. These results allow us to conclude that there is a differential gene regulation related to seizure's occurrence in WAR model, which could explain the susceptibility of this strain to ictal events.

Determining the stability of DNA/RNA hybrid duplexes

Vivianne Basilio Barbosa, Erik de Oliveira Martins, Gerald Weber

Department of Physics, Federal University of Minas Gerais, Belo Horizonte, MG, Brazil

We study the stability of secondary structures in DNA/RNA hybrid duplexes by a nearest neighbor (NN) thermodynamic model. DNA/RNA hybridization has important role in biotechnological applications such as in CRISPR/Cas9 genome editing. Differently from double helices like a DNA/DNA or RNA/RNA which can be self-complementary, sequences of DNA/RNA hybrids constitute a strongly asymmetrical structure due the different compositions the backbone. Nevertheless, it is possible to use a NN model to describe and quantify these interactions for short hybrid sequences with different configurations. Here, we use a set of published experimental sequences of DNA/RNA hybrids and their corresponding melting temperatures measured by UV spectroscopy (Sugimoto et al, Biochemistry, 34, 1995) to determine thermodynamic parameters like free energies formation (ΔG°) and entropic effects (ΔS°) these hybrids. We use a minimization procedure which was recently developed by our group (Weber, Bioinformatics, 31(6), 2015) which bypasses the linear regression of the Van't Hoff plot. This method has the advantage of reducing the uncertainty of the calculated parameters. We also determine the initial free energy factors and an initial entropy factors associated to the formation of the double helix. Such initiation factors are commonly used for DNA/DNA and RNA/RNA, but are currently missing for DNA/RNA hybrids. They are important to supply the necessary energy to form the state of a duplex and are directly associated to the molecular stability of these sequences. Another important factor is the calculation of free energy corrections for terminal base pairs. Our results show that these terminal corrections are very similar to each other for dTrA, dCrG and dGrC base pairs. Finally, we describe physical properties and we discuss the stability of these sequences using the newly calculated parameters. We acknowledge financial support by Capes, CNPq and Fapemig.

Comprehensive profiling and characterization of *Arachis stenosperma* (peanut) and *Meloidogyne arenaria* (plant-root nematode) small-RNAs identified during the course of the infection

Priscila Grynberg, Larissa A. Guimarães, Marcos M. C. Costa, Roberto C. Togawa,
Ana C. M. Brasileiro, Patrícia M. Guimarães

Embrapa Recursos Genéticos e Biotecnologia, Parque Estação Biológica, Brasília-DF

Plant-parasitic nematodes have a worldwide distribution. They are virtually able to infest any human-cultivated plant. Annual losses caused by nematodes on life-sustaining crops are estimated to exceed 14% of the production (approximately 65 billion € of loss worldwide). Previously studies were responsible for major advances in the identification of genes and mechanisms responsible for plants response to the *Meloidogyne*, the root-knot nematode. *Meloidogyne* spp. are obligate endoparasites that maintain a biotrophic relationship with their hosts. During the infection root cells are differentiated into specialized giant feeding cells through the releasing of effector proteins. However, despite the continuing efforts to identify new effectors and plant resistance mechanisms, studies have shown that the repertoire of both systems is limited. Recently, researchers published strong evidence that small RNAs from a phytopathogenic fungus act as effectors. These small RNAs hijack the host RNA interference (RNAi) machinery by binding to *Arabidopsis* Argonaute 1 (AGO1) and selectively silencing host immunity genes. These findings gave new insights on nematode-plant interaction as well as for the development of new control strategies through biotechnological methods. The goal of this work is to verify the possible role of *Meloidogyne arenaria* small RNAs (sRNA) as effectors by identifying, in *Arachis stenosperma* (peanut), downregulated target genes during the infection. *A. stenosperma* plants were infected with approximately 5,000 *M. arenaria* larvae in triplicate. Control and infected *A. stenosperma* roots were collected 3, 6 and 9 days post-infection. The infected samples were pooled. Six samples (3 controls, 3 infected) and two *M. arenaria* J2 small-RNA libraries were sequenced with technical replicates using Illumina HiSeq 2500 system. After adaptor removal, reads were submitted to Infernal 1.1. This program uses Rfam as database. The plant samples output was used as input to test two different plant microRNA prediction tools. For *A. stenosperma* control samples, 325 (65 known) and 1255 (56 known) miRNAs were predicted by miRDP and miR-PREFeR respectively. 34 known and 177 unknown miRNAs were predicted by both tools. For peanut infected samples, 275 (34 known) and 1187 (32 known) miRNAs were predicted by miRDP and miR-PREFeR respectively. 21 known and 157 unknown miRNAs were predicted by both tools. Next steps include: 1) to perform small-RNA prediction at the nematode samples; 2) to classify and validate the plant and nematode small RNAs; 3) to search for nematode small RNA targets in plant; 4) assessment of microRNA differential expression between controls and infected plant samples.

Non-coding RNAs putatively acting as ceRNAs in embryonic stem cells

Raquel Calloni, Diego Bonatto

*Laboratório de Biologia Computacional e Molecular, Centro de Biotecnologia,
Universidade Federal do Rio Grande do Sul*

The cells composing the human body share the same genetic code, but their different transcriptomes, controlled by a complex gene expression regulation system, enable the existence of several different cell types. Recently, a new regulatory mechanism based in the idea that different RNAs can compete for miRNAs ligation was proposed. Named as competing endogenous RNAs (ceRNAs), those molecules share miRNAs response elements with other co-expressed RNAs, acting as miRNAs sponges and leaving the mRNAs targets free to be translated. This competition mechanism is present in several cells, but it has been poorly investigated in embryonic stem cells (ESCs). The aim of this study was search for non-coding RNAs which may act as ceRNAs in ESCs and may be involved in stem state maintenance. For this purpose, RNA-seq data from ESCs and differentiated cells (DIFCs) was downloaded from GEO (GSE64417). The reads were aligned using STAR and differentially expressed mRNAs, lncRNAs, pseudogenes and miRNAs were detected using the package DESeq2. Pearson correlation (PeC) between mRNAs and lncRNAs or pseudogenes and partial correlation (PaC) between mRNAs and ncRNAs controlling for 5 transcription factors (TFs) were estimated using the R packages Hmisc and ggm, respectively. TFs binding the studied genes were retrieved from JASPAR database and gene ontology analysis was performed using the package GOseq. From ESCs upregulated genes ($\log_2FC \geq 1$; $padj < 0.05$), those targeted by ESCs upregulated miRNAs were considered for ceRNA searching. Pairs of mRNAs and lncRNAs or pseudogenes targeted by the same miRNAs and whose PeC was $r \geq 0.6$ ($p < 0.05$) were selected. Positive correlation due to shared regulatory genomic sequences was discarded since the pairs RNAs are coded in different chromosomes. PaC r values pointed that TFs have no influence over the correlations observed. Moreover, pairs observed as positively correlated in DIFCs were removed from the analysis. This filtering process ended up with 107 mRNA-ncRNA pairs. An ontology analysis revealed three mRNAs involved in stem cell population maintenance: FGF2, FZD7 and SALL4. FGF2 is targeted by 3 miRNAs which are putatively sponged by GAS5, POU5F1P3, RP11-L69L16.5 and LINC001194. The following steps are find other pairs putatively important to the ESCs maintenance and include circRNAs in the list of ceRNAs acting in these cells.

Funding support: CNPq

Optimized RNA nearest-neighbor enthalpy and entropy parameters as function of salt concentration

Izabela Ferreira , Elizabeth A. Jolley , Brent M. Znosko , Gerald Weber

Department of Physics, Federal University of Minas Gerais, Brazil

Department of Chemistry, Saint Louis University, United States

Nearest-neighbor (NN) method using enthalpy and entropy parameters is the most common method used to predict melting temperatures for oligonucleotides. The application of those parameters goes well beyond melting temperature prediction, for instance they are crucial for predicting RNA secondary structures. The NN parameters are usually derived from a small set of melting temperatures measured over a range of oligonucleotide concentrations. To obtain the parameters one would typically first obtain the total enthalpies and entropies from a Van't Ho plot linear regression and afterward extract the detailed NN parameters with linear algebra methods. Here, we use a recently developed optimization method which bypasses the linear regression, the NN parameters are instead obtained directly from the melting temperatures. The advantage of this method is that it avoids the increased uncertainty induced by the linear regression of the Van't Ho plot and as a result the predicted temperatures are much closer to the experimental data especially at higher temperatures and long sequences. We apply this technique to obtain the parameters for RNA over a range of different sodium concentrations. The parameter optimization were performed on 18 different RNA sequences at 5 different sodium concentration and 9 to 11 different oligonucleotide concentrations. Our results show that in particular CG-CG NN parameters show very little dependence with sodium concentration. On the other hand NN parameters containing AT base pairs show a quadratic dependence with salt concentration which opens the possibility of deriving salt corrections for enthalpy and entropy parameters specifically for those nearest-neighbor configurations.

Funding: CNPq, Capes, Fapemig, NIH (grant 2R15GM085699)

Sequence-independent metagenomic analysis of animal viromes based on molecular characteristics of small RNAs

Queiroz, L.R., Olmo, R.P., Marques, J.T. and Aguiar, E.R.G.R.

*Department of Biochemistry and Immunology, Instituto de Ciências Biológicas,
Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil*

Viruses are obligatory intracellular parasites that require the host machinery to replicate. Since viral RNAs normally present different molecular characteristics from those present in cellular RNAs, viral RNA intermediates can be recognized and degraded by different antiviral mechanisms that include RNA interference pathways, miRNA, siRNA and piRNA. These virus-derived small RNAs are imprinted with molecular characteristics that reflect their origin. Thus, since molecular features of small RNAs such as size, polarity and base preferences depend on the type of viral substrate and host mechanism of degradation, the pattern of small RNAs generated in infected cells can be used as a molecular footprint to identify and characterize viruses independent on sequence homology searches against known references. In this work, we analyzed 27 small RNA libraries from a broad range of organisms infected with viruses, including plants, arthropods and mammals, to determine how molecular characteristics of small RNAs could be applied to identify viral and non-viral sequences. We observed that production of small RNAs ranging from ~20 to 23 nt symmetrically from both strands, typically associated with the activation of siRNA pathway, was a conserved and specific response to viral infection in different organisms. Additionally, in arthropods, we noticed that other small RNA sources such as Endogenous Viral Elements (EVEs) and Transposable Elements (TEs), showed profiles more consistent with the activation of the piRNA pathway (e.g. small RNAs ranging from 24 to 30 nt, asymmetrical in polarity, and a strong 5' U preference). These molecular characteristics of small RNAs allowed us to use Hierarchical Clustering based on Pearson correlation to classify sequences independently of homology searches against reference databases. Our results indicated that small RNA patterns are able to separate clusters of sequences containing viruses, EVEs and transposable elements. Using pattern-based analysis we successfully differentiate 10 viral sequences from 76 EVEs and 8311 TEs in six different organisms. This strategy can help overcome a great limitation of virus discovery by metagenomic strategies, since it does not require sequence similarity searches against known references. We are currently expanding our strategy to include the analysis of the ORF profile along assembled contigs and the use of di- and tri-nucleotide frequencies to identify and classify viral sequences.

Financial Support: CNPq, CAPES and FAPEMIG

LncRNAPlant-Finder: a tool for prediction of long non coding RNAs in plants

Tatianne da Costa Negri¹, Pedro Henrique Bugatti¹, Priscila Tiemi Maeda Saito¹, Douglas Silva Domingues^{1,2} e Alexandre Rossi Paschoal¹

¹Programa de Pós-Graduação em Bioinformática - PPGBIOINFO, Universidade Tecnológica Federal do Paraná, Campus Cornélio Procópio; ²Departamento de Botânica, Instituto de Biociências, Universidade Estadual Paulista, Campus de Rio Claro

* Corresponding author: paschoal@utfpr.edu.br

Long non coding RNAs (LncRNAs) correspond to a eukaryotic non-coding RNA class with more than 200 nucleotides in length. They have emerging attention in the last years as a potential layer of gene expression in cells. However, LncRNAs mechanisms in plants are still poorly known. Moreover, there is a lack of specific computational approaches for lncRNA prediction in plants, considering that the biological mechanism of this ncRNA class is different from mammals, which there are several tools for prediction. Having this in mind, we present the LncRNAPlant-Finder, an approach for lncRNA identification in plants. To build this tool, we used publicly available lncRNA and transcript (mRNA) sequences from six plant genomes: *Arabidopsis thaliana*, *Cucumis sativus*, *Glycine max*, *Oryza sativa*, *Populus trichocarpa* and *Setaria italica*. All the data was extracted from the public databases PLNlncRbase, GREENC and Phytozome, where we used 22,543 lncRNAs and 29,960 transcripts. We applied pattern recognition techniques in a total of 85 features based on sequence and structure from lncRNAs and transcripts (e.g. GC content, ORF, dinucleotide and trinucleotide distribution) in order to select the best features for classification. Sequences were also processed using: (i)- CD-Hit-EST: to avoid sequence redundancy; (ii)- txCDSPredict: for ORF prediction; (iii)- in-house PERL scripts to calculate di and tri-nucleotides frequency, GC context, normalization and generating an ARFF file. All feature selection and classification processes were done using Weka 3.8.0. We detected 16 best features for classification after feature the selection process. These features were used to compare six classification methods. The J48 method obtained the best results with: (i)- Correctly Classified Instances (CCI), ≈97%; (ii)- Incorrectly Classified Instances (ICI), ≈ 3%; (iii)- Correct lncRNAs (CL), 22.021 (≈97,7%); (iv)- Correct Transcripts (CT), 28.812 (≈96,25%); (v)- Error lncRNA (EL), 522 (≈2,3%); (vi)- Error Transcripts (ET), 1.148 (≈3,9%). These results point out a promising approach to help lncRNA identification in plant genomes.

Support: CNPq (#454505/2014-0); Fundação Araucária: N° 019/2015.

Transcriptional memory in coffee response to drought

Fernanda Alves de Freitas Guedes¹, Priscilla Nobres¹, Daniela Cristina Rodrigues Ferreira¹, Régis Lopes Correa¹, Marcelo Ribeiro Alves², Fábio Murilo DaMatta³, Márcio Alves-Ferreira¹

¹*Federal University of Rio de Janeiro (UFRJ);* ²*Oswaldo Cruz Institute (FIOCRUZ-RJ);*

³*Federal University of Viçosa (UFV)*

Harsh environmental conditions can induce different plant stress responses for which a cross-talk may occur. Water deprivation is an important limiting to crop productivity. Modulation of cellular pathways triggered by abscisic acid (ABA) probably involving receptor-like kinases is a crucial step for drought response. The Reactive Oxygen Species also participate in drought response but the toxic increase of their levels induced by the stress must be mitigated to protect cells against oxidative damages. The complex expression regulation of genes involved in plant drought response counts not only on transcription factors (TFs) but also on transcriptional memory. Here, RNA-Seq approach was used to investigate drought responses of *Coffea canephora* clone 109 and 120, respectively sensitive and tolerant to drought. Illumina sequencing allowed us to identify 826 differentially expressed genes (DEG) in the tolerant clone and 135 in the sensitive clone. "Response to ABA" and "heat acclimation" GO categories were exclusively enriched in tolerant clone DEG, respectively after one and three drought cycles. Coffee genes that exhibited altered expression after the first and the third drought exposures were considered memory genes. For tolerant clone, 49 genes exhibited transcriptional memory after multiple drought exposures. Three memory receptor-like kinases probably related to ABA signaling were found to interact with the Cc02_g02350 heat shock protein whose memory profile was confirmed by qPCR. Small RNA profiling (sRNA-Seq) data were analyzed using two different softwares and the miRBase database in order to identify microRNAs in coffee leaves. Conserved regulatory miRNAs and their putative targets were identified, together with putative novel miRNAs. While tolerant plants acclimate to stress, multiple drought exposures seems to induce oxidative stress in the sensitive clone which, in turn, may lead to induction of programmed cell death. Our findings show that transcriptional memory modulates expression of drought-responsive genes and contributes to drought tolerance in *C. canephora*.

Financial support: FAPERJ, CAPES, CNPq.

Mirtrons: computational feature analysis and miRNA comparison

Tamires Priscila da Costa¹, Douglas Silva Domingues^{2,3}, Alexandre Rossi Paschoal^{1,2}

¹ Departamento de Computação – DACOM, Universidade Tecnológica Federal do Paraná, Cornélio Procópio, ² Programa de Pós-Graduação em Bioinformática - PPGBIOINFO, Universidade Tecnológica Federal do Paraná, Cornélio Procópio and ³ Departamento de Botânica, Instituto de Biociências, Universidade Estadual Paulista, Rio Claro

The advances in genome sequencing technologies, as well as in bioinformatics approaches allow numerous scientific analyses using publicly available data. Among these data are the microRNAs (miRNAs). They correspond to one class of non-coding RNAs that acts as a post-transcriptional regulator of mRNA level in cells. Recently, a novel microRNA class, named mirtrons, was identified in several model organisms and they have different biogenesis when compared to canonical miRNAs. Mirtrons are processed by splicing step instead of the cleavage by a specific enzyme. There are more than 140 databases specific for noncoding research, mostly devoted to miRNA data. However, the biological computational view of mirtrons is not sufficiently addressed. In this study, we did a computational features analysis on mirtron public data and compared against canonical miRNAs. Our interest is to identify mirtron-specific properties that could be applied in future *in silico* applications. Mirtron data was extracted from literature, using public curated data of *A. thaliana*, rice, human, mouse, fly and worm genomes. We analyzed five features based on sequence and structure aspects: GC content, length, sequence nucleotides distribution (up to K-mer <4), free energy and conservation. These analyses were performed using in-house PERL scripts, R studio and BLAST program for alignment. We have found differences in the free energy distribution, mainly in plant mirtrons. Specifically for conservation analysis we focused on plant and fly genomes. Our results suggest conservation in flies genomes while we did not find any significant results in plants. Overall, this work provides a feature overview on mirtrons literature data. We believe that this report will contribute to future research in computational approaches to understand mirtron distribution and characteristics.

Support: CNPq (#454505/2014-0); PIBIC/UTFPR (#02/2015)

E ect of Cy3 and Cy5 dyes on the hydrogen bonds of oligonucleotides

Pâmella Miranda, Luciana M. Oliveira, Gerald Weber

Department of Physics, Federal University of Minas Gerais, Brazil

Cyanine dyes Cy3 and Cy5 exhibit intense visible fluorescence and when attached to the 5' ribose terminus of DNA stabilize the duplex. Oligonucleotides modified with cyanines are widely used in biotechnological techniques such as microarrays and real-time PCR. They also allow for nanoscale measurements by using Forster resonance energy transfer (FRET). Despite its importance little is known on how these dyes affect the intramolecular interactions of the base pairs they are attached to. A detailed knowledge of these interactions would allow for a better understanding and perhaps would lead to the optimization of experimental techniques. Here we use the Peyrard-Bishop (PB) mesoscopic model to obtain estimates of the hydrogen bonds and stacking parameters of DNA with attached Cy3 and Cy5 molecules. The PB model has the unique ability to consider separately the hydrogen bonds and the stacking interactions by using independent potentials. We applied the PB model successfully to predict the molecular interactions in DNA and RNA. In particular, we confirmed independent NMR measurements on AU in RNA showing a stronger hydrogen bond than AT in DNA. More recently we were able to correctly predict single hydrogen bonding in specific GU tandem base pairs. We also showed how the solvent affects the hydrogen bonds of terminal base pairs. Here, we use a recently published melting temperature data set of 35 DNA sequences with cyanine dyes and 10 control sequences. By combining the PB model and adjusting the parameters to provide a better fit to the experimental melting temperatures we are able to determine not only the interactions of the dyes with the DNA terminus but also how those affect the neighboring base pairs. In particular we show an enhanced thermal stability induced by the cyanine dyes. The method also allows a detailed map of the opening probabilities along the oligonucleotide sequence where we show how the dyes affect the stability at the base pair level.

Funding: CNPq, Capes, Fapemig

SNP discovery in the *Klebsiella pneumoniae* transcriptome after polymyxin B induction in combination with abiotic stresses using RNA-Seq technology

¹Thiago Cardoso Pereira Carneiro¹, Guilherme Loss de Moraes², Gisele Lucchetti da Silva, Guadalupe del Rosario Quispe Saji², Ana Tereza Ribeiro de Vasconcelos², Marisa Fabiana Nicolás²

¹Universidade Estácio de Sá, ²The National Laboratory for Scientific Computing LNCC. Av. Getúlio Vargas, 333 Quitandinha, 25651075, Petrópolis, RJ Brazil.

The Single Nucleotide Polymorphism (SNP) can be characterized as the sequence variation observed in an individual position in the genome, and this may involve direct and relevant way in the formation of a protein. SNPs can be used to type bacterial strains and also used to guide site-specific mutation studies. *In silico* detection of SNPs are currently applied for bacterial genomics studies. *Klebsiella pneumoniae* is a Gram-negative, rod-shaped bacterium frequently associated with nosocomial and community-acquired infections. The emergence and subsequent global spread of strains producing *Klebsiella pneumoniae* carbapenemase (KPC) represents a significant threat to public health. Many KPC infections can only be treated when resorting to last-line drugs such as colistin and polymyxin B (PB). However, resistance to these antibiotics is also observed, although insufficient information is yet described on its mode of action as well as with combination of abiotic stresses. In a recent study conducted in *Pseudomonas aeruginosa* was reported that free iron in the culture medium increases mutagenesis in the presence of cationic antimicrobial peptides (cAMPs). Considering that PB is a type of cAMP and in view of this information, the goal of this study was to investigate the SNPs induced by the effect of PB in combinations with abiotic stresses, such as high level concentration of Fe or low pH. The cDNA data of *K. pneumoniae* subsp. *pneumoniae* Kp13 were generated by illumina platform Hiseq (www.illumina.com) in induced conditions (PB) and high concentrations of Fe or low pH. Firstly, the RNA-seq data were trimmed and then mapped against the Kp13 genome reference (NCBI Bioproject PRJNA78291) by the bowtie2 (bowtie-bio.sourceforge.net/). Then, the result files were ordered and indexed by Samtools (samtools.sourceforge.net/). The SNP search was done with GATK (software.broadinstitute.org/gatk) and the IGV (software.broadinstitute.org/software/igv) allowed a better visualization and analysis of these mapped SNPs. Finally, SNPs were classified according to synonymous or non-synonymous mutations. In the case of non-synonymous mutations, it was investigated the CDS role in the bacterial response to antimicrobial stresses. Comparing PB versus PB+pH or PB+Fe we observed a higher incidence of SNPs in PB+Fe condition. An interesting gene containing three non-synonymous mutations as well as one stop loss mutation was *sodB*, which encodes for a superoxide dismutase Fe (SOD-Fe), which is a major enzyme produced by microorganisms to evade the potentially damaging reactive oxygen species (ROS) during the stress imposed by antimicrobials.

Integrative analysis of transcriptomics and metabolomics data: adaptation of *Propionibacterium freudenreichii* to long-term survival in nutritional shortage

Flavia Figueira Aburjaile^{1,2,3,4}, Anderson Miyoshi⁵, Artur Silva⁴, Vasco Azevedo¹, Yves Le Loir^{2,3} and Hélène Falentin^{2,3}

¹Laboratory of Cellular and Molecular Genetics, UFMG, Minas Gerais, Brazil; ²INRA, UMR 1253, Science et Technologie du Lait et de l’Oeuf, Rennes, France; ³Agrocampus Ouest, UMR1253, UMR Science et Technologie du Lait et de l’Œuf, Rennes, France;

⁴Center of Genomics and System Biology, UFGPA, Pará, Brazil; ⁵TecnoGen, Laboratório de Tecnologia Genética, UFMG, Minas Gerais, Brazil

Propionibacterium freudenreichii is a bacterial species belonging to Actinobacteria phylum, known for its survival in long periods under adverse environmental conditions. This bacterium is widely used in dairy industry for the ripening process of Swiss-Type cheeses, such as Emmental and also presents potent strain-specific probiotic effects. In this study, *P. freudenreichii* CIRM-BIA 138 was grown for 11 days in a culture medium with nutritional shortage. Bacterial survival rate was assessed by optical density and CFU counting. Gene expression and biochemical analysis were also conducted to investigate the bacterial adaptive response to nutritional shortage condition. This strain maintained a high population level of 10^8 CFU/ml for the entire period. The available carbon and free amino acids sources and organic acid produced by the strain were monitored in the bacterial supernatant throughout survival for 11 days. Lactate was the first carbon source for *P. freudenreichii* CIRM-BIA 138 to be exhausted in the medium. RNA-seq analysis demonstrated different metabolic behaviors across the conditions of stationary entry phase and exponential phase. At the beginning of stationary phase, *P. freudenreichii* CIRM-BIA 138 dramatically reduces several metabolic processes as glycolysis, oxidative phosphorylation and Wood-Werkman Cycle, seeking an alternative metabolism facing pathways that promote new energy sources, such as carbon and nitrogen. Moreover, it is noted that, the processes of transcription, translation and secretion of proteins are decreased at the stationary growth phase. The transcriptomic and metabolomic data integration allowed us to deduct the strategies involved in adaptation, persistence and long-term survival of *P. freudenreichii* in nutritional shortage.

Funding: CAPES/COFECUB and CNPq.

Single nucleotide variation analysis in microRNA target regions in colorectal cancer

Jéssica Blanco, Natasha Andressa Nogueira Jorge, Fabio Passetti

Laboratório de Genômica Funcional e Bioinformática, Instituto Oswaldo Cruz, Fiocruz

Colorectal cancer is the third most common cancer in the world. In 2012, The Cancer Genome Atlas Network Consortium reported that different colorectal tumor samples have distinct genetic alterations. MicroRNA (miRNA) is a key player in the control of eukaryotic gene expression. This family consists of small noncoding RNAs that prevent translation of target messenger RNAs (mRNAs), thus reducing the expression of the gene encoded in the mRNA. The region responsible for the miRNA target recognition is termed seed. Single nucleotide polymorphism (SNP) is commonly found in different types of cancer and this alteration is characterized by a change of a single nucleotide at a particular position in the genome. SNPs have been described in several regions associated with miRNAs in tumor samples, such as seed regions. This project intends to evaluate, qualitatively and quantitatively, the frequency of SNPs in miRNAs target regions and their adjacent genomic regions in colorectal cancer samples. Small RNA high throughput sequencing data of 5 rectal tumor samples and matched normal tissue were obtained in TCGA database (downloaded: 01/2016). SNP call was performed using the VARSCAN2 and information concerning the target regions of miRNAs were obtained in the TargetScan database. We investigated SNP effect in four different regions: seed region, miRNA pairing region outside seed region, 200 nucleotides upstream target region and 200 nucleotides downstream target region. SNPs in several genes were found common to all patients analyzed, as follows: 17 genes presented alterations in the target seed region, while 407 genes in the region 200 nucleotides upstream, 39 genes in the microRNA region outside seed and 384 genes in the region 200 nucleotides downstream. These preliminary results motivate us to expand this analysis to other 25 paired colon adenocarcinoma samples; we believe this project will help improve our understanding of the changes in microRNA regulation and its role in colorectal cancer. The effects of these modifications still need further study so that it can be possibly used in the future in therapy and cancer prevention.

Financial Support: CNPq, CAPES, Fiocruz, FAPERJ.

Transcriptome analysis of high-temperature stress in yeast during industrial scale bioethanol production

Luciana Souto Mofatto¹, Osmar Vaz de Carvalho-Netto^{1,2}, Gleidson Silva

Teixeira^{1,2}, Silvio Roberto Andrietta³, Maria da Graça Stupiello Andrietta³,

Gonçalo Amarante Guimarães Pereira^{1,2}, Marcelo Falsarella Carazzolle¹

(1) Laboratório de Genômica e Expressão, Instituto de Biologia, Universidade Estadual de Campinas; (2) GranBio – BioCelere; (3) Centro Pluridisciplinar de Pesquisas Químicas Biológicas e Agrícolas (CPQBA), Universidade Estadual de Campinas.

Currently there is a growing demand for renewable energy sources, since the traditional energy sources are becoming limited, such as petroleum oil. The production of bioethanol, an alternative energy source used in Brazil, is based on the fermentation of sucrose from sugarcane feedstock using highly adapted industrial strains of the yeast *Saccharomyces cerevisiae*. In the industrial environment, yeasts are usually under several stress conditions, such as high temperature, low pH, bacterial contamination and others. The high-temperature tolerance is a desirable phenotype because can decrease production costs and increase the productivity in the bioethanol industry. In this context, bioinformatics has a key role by enabling large-scale analysis of transcriptome in industrial strains, identifying genetic aspects involved in the stress tolerance process of these yeasts. Therefore, the aim of this study was identifying the transcriptomic profile of industrial yeasts (*Pedra II - PEII*) under two temperature conditions in industrial environment operating in fed batch fermentation. The samples were collected in triplicates direct from the industrial fermentation tanks after 4 and 8 hours of fermentation under 32°C (control group) and 38°C (test group). mRNA samples from yeasts were sequenced using Illumina HiSeq2000 and composition of fermentation broth was measure by HPLC. Reads from each RNA-seq library were aligned against a reference gene database constituted by all *Saccharomyces cerevisiae* S288c genes and 20 *PEII*-specific genes. In-house perl scripts were performed to calculate the number of aligned reads per gene (read counts) and gene expression levels (RPKM). For differentially expressed genes (DEG) analysis, read counts values were submitted to negative binomial statistical test using R/Bioconductor packages DESeq2 and edgeR and filtered by p-value < 0.05, |fold-change| >= 2 and RPKM >= 1. The DEG were submitted for Gene ontology (GO) enrichment analysis using SGD database (www.yeastgenome.org). As results, 447 DEG were found (181 up-regulated and 266 down-regulated) after 4 hours of fermentation and 1080 DEG (540 up-regulated and 540 down-regulated) after 8 hours. GO terms as response to heat, stress, and temperature stimulus were enriched in up-regulated genes on both fermentation time, but cell wall organization and protein folding processes appeared only after 4 hours. For down-regulated genes, ergosterol, sterol and lipid biosynthetic process were enriched after 4 hours while oxidative phosphorylation and aerobic respiration processes appeared after 8 hours of fermentation. Finally, a protein-protein interaction network of DEG were constructed for identifying hub proteins that can represent master regulators of this process.

The assessment of the impact of small deletions within human protein domains using transcriptome data: a preliminary analysis.

Fernanda Oliveira, Gabriel Wajnberg, Fabio Passetti

*Laboratory of Functional Genomics and Bioinformatics, Oswaldo Cruz Institute,
FIOCRUZ, Rio de Janeiro, Brazil.*

Deletions is an example of sequence polymorphism, which can alter the encoded protein sequence. These alterations within the amino acid sequence can be associated to many diseases in human, such as cancer. High throughput sequencing data can be used to identify novel small polymorphisms, such as deletions. This work aims the analysis of the impact of deletions within protein coding domains associated with tumors, with the use of transcriptome data generated by RNA-Seq. Our preliminary analysis identified coding protein domains affected by small deletions, up to 100 nucleotides in length, in RNA-Seq data available in the public database SRA from matched normal and tumor samples in six lung cancer patients. We identified 734 unique affected protein domains: 123 only in normal samples, 71 exclusively to tumor samples and 540 found affected in both normal and tumor samples. Deletions were detected in protein domains with high probability of being associated with cancer biology, such as deletions occurring in protein tyrosine kinase domains in FLT4 and HTR3A genes in both normal and tumor samples. This type of protein domain is frequently affected by mutations in cancer, including lung cancer. We also identified the zinc finger protein domain C2H2 affected in five different genes only in tumor samples, such as: H1NFP, RBAK, ZNF468, ZNF234 and ZNF571. These genes have different functions, but all of them are associated in the regulation of transcription interacting with transcription factors and other components such as E2F1 transcription factor (RBAK gene function). In conclusion, our preliminary results shows that these small deletions detected using transcriptome data of 6 lung cancer patients may be disrupting some important protein domains associated with cancer biology and we can contribute to identify novel cancer genetic markers with this data.

Financial Support: CAPES, FIOCRUZ, FAPERJ and CNPq.

Non-coding RNAs in the genus *Aeromonas*

Jean Carlos Machado da Costa¹, Alexandre Rossi Paschoal^{1,2}, Cynthia Maria Teles Fadel Picheth³, Fabio de Oliveira Pedrosa⁴, Maria Berenice Reynaud Steffens^{1,4}

¹*Graduate Program in Bioinformatics, 2Department of Bioinformatics in UTFPR,*

³*Department of Medical Pathology, 4Department of Biochemistry and Molecular Biology*

The *Aeromonas* spp are Gram-negative bacteria that do not form spores, they are chemoorganotrophic and facultative anaerobes. The first genre description dates from the late nineteenth century, however the genre was only set after a phylogenetic analysis of 16S rRNA in 1992, establishing a new *Aeromonas* genus, which comprises the *Aeromonadaceae* family, *Aeromonadales* order, Proteobacteria class and subclass gama-Proteobacteria. This bacterium has a high degree of pathogenicity causing opportunistic infections. In Bacteria, non-coding RNAs with regulatory function (ncRNAs) can modulate physiological responses and act by different mechanisms such as RNA-RNA pairing bases and RNA-protein interactions. Technologies for non-coding RNA prediction analysis such as the Infernal program (Inference RNA Alignment), can predict different types of ncRNAs, indicating that the amount of regulatory ncRNAs can be higher than previously thought. Traditionally, these approaches along with the help of targeting programs as TargetRNA2 and more advanced technologies, such as RNA-Seq, allows to identify ncRNAs involved with the virulence process, biofilm formation, resistance to antibiotics and survival. The aim of this study is to identify ncRNAs present in the *Aeromonas* strains: *A. hydrophila*, *A. caviae*, *A. sobria*, *A. trota* and *A. veronii*. The genomes deposited inside the NCBI database were used as input data for the Infernal 1.1 tool. The output data were separated into ribosomal RNAs, carrier RNAs and ncRNAs. The ncRNAs were classified as smallRNAs, regulatory ncRNAs and riboswitches. The identity of ncRNAs was determined on the Rfam database. The TargetRNA2 software was used for target prediction. A total of 237 ncRNAs were found. Fifty ncRNAs were assigned as regulatory, 6 as riboswitches, 6 as microRNAs and 175 as smallRNAs. We identified regulatory ncRNAs that acts in cis, ncRNAs that associates with Hfq, ncRNAs involved in virulence, pathogenicity, responsible for biofilm formation and cell survival. Regarding microRNAs found in this analysis in prokaryotes, we have the following hypothesis: the isolates came from clinical specimens, which strengthens a pathogen-host relationship and the predictor program identifies the sequence as microRNA since this sequence is in your database. They were compared with other enteropathogenic bacterium references (*E. coli* and *Salmonella* spp.), showing a great similarity between these species.

Altered gene expression by control unspecific dsRNAs: an inquiry

Sandra Grossi Gava^{1,2}; Naiara Cristina Clemente Santos Tavares de Paula¹; Anna Christina de Matos Salim³; Flávio Marcos Gomes Araújo³; Guilherme Oliveira⁴, Marina de Moraes Mourão¹

¹Grupo de Helmintologia e Malacologia Médica, CPqRR; ²Instituto de Ciências Biológicas, UFMG; ³Plataforma de Sequenciamento Genômico e transcriptoma NGS, CPqRR; ⁴Instituto Tecnológico Vale, ITV

RNA interference is long-serving and still the only reverse genetic tool available for gene function studies in trematodes. Most of de RNAi assays in trematodes have been performed in *Schistosoma*, especially in *S. mansoni*. Gene silencing generally use a nonrelevant dsRNA from another species as controls and quantitative real time PCR (qPCR) to measure the knockdown levels achieved. Despite the applicability of RNAi to study many genes in schistosomes and others helminthes parasites, several authors have noticed inconsistencies associated with this technique. To globally check if there are genes affected by unspecific dsRNA exposure, schistosomula (~500,000 larvae) were cultivated and exposed to 100 nM of unspecific dsRNA synthetized from the Green Fluorescent Protein (GFP) or mCherry, two sequences with no similarity with *Schistosoma* genome and widely used by the scientific community. After two days of culture, total RNA extraction was carried. RNA-Seq libraries were prepared according to the *Truseq stranded mRNA Library Prep* protocols and were sequenced on *Illumina HiSeq 2500* platform. We generated 10 paired-end libraries containing reads of 100 bp, ranging from 34 to 92 million reads per library, with GC content of 38-39%. The sequences were aligned to the *S. mansoni* reference genome using STAR with more than 87% of uniquely mapped reads. Counts of reads aligned were obtained with the sub-command multicov in BEDTools suite. To ascertain genes differentially regulated (DEGs) after non-specific dsRNA exposure, we compared the expression profiles with the untreated controls. RNA-seq analysis resulted in 6 DEGs in the GFP dsRNA treatment and 3 DEGs in the mCherry dsRNA treatment (edgeR *P*-value < 0.01, FDR < 0.05). KEGG and GO databases were used to elucidate the functional classifications of these DEGs. The majority of DEGs coded for uncharacterized proteins and none of them are used as control in RT-qPCR experiments. We visualize the sample-to-sample distances in a heatmap and PCA analysis and found that the biological replicates are more prone to cluster than unspecific and untreated controls. Here, we conclude that there are more differences between biological replicates, than due to the treatment with GFP or mCherry unspecific dsRNAs. These observations may be relevant to other model systems applying RNA interference for gene function assessment.

Financial support: CAPES; FAPEMIG; CNPq, CPqRR-FIOCRUZ, European Comission 7th Framework, A-ParaDDisE

PIWI-interacting RNA and small nucleolar RNA signatures of smokers and non-smokers in lung adenocarcinoma

Natasha Andressa Nogueira Jorge¹, Gabriel Wajnberg¹, Benílton Carvalho², Carlos G. Ferreira³, Fabio Passetti¹

¹*Laboratório de Genômica Funcional e Bioinformática, Instituto Oswaldo Cruz, Fundação Oswaldo Cruz, Rio de Janeiro, RJ, Brazil;* ²*Department of Medical Genetics, School of Medical Sciences, State University of Campinas, Campinas, SP, Brazil;* ³*Clinical Research Coordination, Instituto Nacional de Câncer, Rio de Janeiro, RJ, Brazil*

Lung cancer is one of the most frequent type of cancer worldwide. However, the majority of the cases are diagnosed when the disease is clinically advanced, leading to poor prognosis. PIWI-interacting RNAs (piRNAs) and small nucleolar RNAs (snoRNA) are two classes of small non-coding RNAs which expression was reported altered in several types of cancer. We investigated differentially and constitutively expressed piRNAs and snoRNAs in publicly available smoker and non-smoker lung adenocarcinoma small high throughput sequencing data. Eight non-smoker and twelve smoker matched tumor and control samples were aligned on the human genome using Novoalign. The differentially expressed piRNAs and snoRNA were obtained using EdgeR and the constitutively expressed were assigned using variance analysis. We identified several distinct sets of differentially expressed piRNAs and snoRNAs among the samples studied. SNORD89 was found less expressed in tumor samples and more expressed in smoker's samples, while piR_023057 and piR_015341 were found more expressed in non-smokers. Most of the snoRNA were also reported as anomaly expressed in cancer. We also identified snoRNA and piRNA constitutively expressed in both tumor and control samples, regardless of smoke status, including SNORD43 and SNORD57. The expression of SNORD43 was already reported as stable and useful as normalization parameter for prostate, bladder and renal cancer. Although none of our piRNA has been previously reported in cancer samples, PIWI proteins have been reported altered in lung cancer and related to patient survival. Our findings stress the importance of non-coding RNAs in cancer biology and prognosis and their increasing role as potential biomarkers for diagnosis and treatment. In conclusion, we identified a set of piRNA and snoRNA that can distinguish smokers from non-smokers in lung cancer.

Financial Support: CNPq, CAPES, Fiocruz, FAPERJ.

De novo transcriptome assembly of the extremophile plant *Calotropis procera*

Rivas, Rebeca¹; Bezerra-Neto, João Pacífico²; Pandolfi, Valesca²; Coêlho, Maria Reis Velois¹; Santos, Mauro Guida¹; Benko-Iseppon, Ana Maria²

¹Universidade Federal de Pernambuco, Departamento de Botânica, Laboratório de Fisiologia Vegetal, Recife, PE, Brazil; ²Universidade Federal de Pernambuco, Departamento de Genética, Laboratório Genética e Biotecnologia Vegetal, Recife, PE, Brazil

Calotropis procera (Apocynaceae) is an evergreen shrub found in arid and semiarid environments, whose parts (shoot, leaf, root, flower and especially latex) are widely used in phytotherapy. Its anti-inflammatory properties make this species an excellent candidate for the search of antimicrobial peptides (AMPs). The objective of this work was to analyse the *C. procera* transcriptome under environmental stress and to identify AMPs, focusing on the thaumatin PR-5 family. High throughput sequencing was performed using Illumina Hi-seq 2500 2 x 100 bp reads, from six libraries of the *C. procera*, treated (30min, 2h, 8h and 45 days under 100 mM NaCl stress imposition) and untreated (0h and 45 days, controls). The sequencing data analysis and assembly were performed using Trinity platform. Assembled sequences were compared against the Universal Protein Resource (UNIPROT). To evaluate the AMPs, these sequences were submitted to the tool AMP-Identifier v.1.0. We obtained 284 million reads including 26, 29, 62, 26, 46 and 95 million for 0 h, 30 min, 2 h, 8 h, and 45 days after salinity stress and 45 days control, respectively. De novo assembled reads generated 224,652 transcripts with a mean of 745 bp in length (224 and 31,249 bp for minimum and maximum, respectively), N50 of 2,525 bp and GC content 39.33%. The transcripts comprise 134,461 unigenes with a mean length of 417 bp and N50 of 1,606 bp. After sequence annotation (UNIPROT database), 80.64% (181,149 transcripts) were predicted and 19.36% (43,503 transcripts) were annotated as unknown. The expression analysis showed 50% of the transcripts with fold change (FC) < 2 for each comparison. FC > 2 and < 10 were represented by 25% and 37% of up and down-regulated transcripts, respectively. In the search for AMPs, via AMP-Identifier 880 candidates were obtained, distributed in 40 AMPs families. After AMP prediction (CAMP_{R3}) and conserved domain search, we obtained 95 AMPs distributed in six families: cecropin (40), cystatin (2), defensin (7), moricin (2), thaumatin (42) and transferrin (2). All 42 possible thaumatins presented the complete expected domain, 32 (76%) thaumatin candidates presented signal peptide and 40 (95%) presented 16 cysteines conserved with eight disulphide bonds, typical of thaumatins. After expression analysis, nine thaumatins were upregulated (FC > 2) in early times of salinity (30 min, 2 h, and 8 h) and two were down-regulated (FC < -2) in the late times after stress (45 days). Thaumatin expression data indicate that increased expression occurs in the first hours after stress in response to salinity.

Financial support: CAPES/CNPq.

TFBS prediction in sugarcane using binding sites prediction from PlantTFDB server

Mauro de Medeiros, Danielle Izilda Rodrigues da Silva, Alan Durham, Glaucia Souza Mendes

Ph.D student in Bioinformatics IME/USP, Ph.D student in International Cellular and Molecular Plant Biology ESALQ/USP, Associate Professor IME/USP and Associate Professor IQ/USP.

Brazil is the world's largest sugarcane producer with productivity close to 70 or 80 tons per hectare [t/ha]. However, such productivity is low if compared with its maximum yield (around 380-472 t/ha). Gene regulation is considered the primary mechanism for activating all biological potential. Thus, the search for new varieties that express higher productivity can begin by understanding gene regulation processes. However, due to the complexity and diversity of the gene regulation mechanisms, research efforts have limited the study mainly on the promoter region. This region is composed of two sets. The first set is the core promoter - DNA sequence of about 100 nucleotides (nt) and the subsequences TATA box, Inr and TSS (transcription start site). The second set consists of DNA sequences that are upstream of the core promoter subsequences such as the CAAT box and TFBS (Transcription Factor Binding Site). The discovery of these regions can be achieved through *in vivo* experiments such as Chromatin Immunoprecipitation followed by Sequencing (ChIP-Seq), systematic evolution of ligands by exponential enrichment (SELEX) and DNase I hypersensitive mapping. However, these tests can be too expensive and are not always suitable for non-model organisms, like sugarcane. To address this challenge some studies have been using the *in silico* approach. The aim of this work was to search TFBSs - described in the literature - in the promoter region of differentially expressed genes of sugarcane samples subjected to water stress. To solve this task we used the FIMO tools and the PlantTFDB database binding sites. The FIMO application performs the individual search for motifs through Position-Specific Scoring Matrix - PSSM. The PlantTFDB is a non-redundant collection of 674 TFBSs distributed in 156 plant species. To task discovery, we used 6 groups from 100 DNA sequences of the promoter region of differentially expressed genes. According to our analysis, in all groups 25% to 30% of the evaluated sequences were classified as TATA-box promoter region. However, when the assessment was related to the number of TFBS from Transcription Factors responsive to drought, Promoter groups from stressed leaf samples were at least twice as large as its counterpart. Moreover, this scenario is not the same when compared with dry and irrigated root. This difference between the two organs may indicate that each tissue has its own mechanism of regulation.

Integrative bioinformatics data analysis of Nile Tilapia microRNAs

Luiz Augusto Bovolenta^a, Danillo Pinhal^b, Simon Moxon^c, Arthur Casulli de Oliveira^a, Pedro Gabriel Nachtigall^b, Marcio Luis Acencio^c, Cesar Martins^a and Ney Lemke^a

^aDepartment of Physics and Biophysics, Institute of Biosciences of Botucatu, São Paulo State University - UNESP, Brazil; ^bDepartment of Genetics, Institute of Biosciences of Botucatu, São Paulo State University - UNESP, Brazil; ^cSchool of Biological Sciences, University of East Anglia, Norwich Research Park, Norwich, United Kingdom; ^dDepartment of Cancer Research and Molecular Medicine, Faculty of Medicine, Norwegian University of Science and Technology - NTNU, Trondheim, Norway; ^eDepartment of Morphology, Institute of Biosciences of Botucatu, São Paulo State University - UNESP, Brazil.

MicroRNAs (miRNAs) are considered essential regulators of several biological processes, such as development, body axis patterning, immune responses, cell fate, proliferation and death. MiRNAs are small RNA molecules that contain approximately 22 nucleotides that are originated from longer hairpin-like RNA molecules (pri-miRNAs). These small RNA molecules silence gene expression either via mRNA degradation or preventing mRNA from being translated. In teleosts, miRNAs have been mainly associated with sexual differentiation, diseases markers, developmental and growth regulatory elements and environmental modulation effects. The Nile tilapia (*Oreochromis niloticus*), an teleostei species, has been globally recognized as a commercially valuable fish due to its phenotype characteristics, such as high endurance, easy adaptability, good consumer acceptance and rapid growth in a variety of aquaculture systems. In this sense, miRNA data are valuable to unravel molecular mechanisms responsible for phenotypes with economic value, specifically how miRNAs could enhance Nile tilapia productivity. To identify known and novel miRNAs associated with interesting phenotypes in Nile tilapia (sex differentiation, muscle growth and disease resistance), we performed an integrative bioinformatics analysis in attempt to unravel biological mechanisms of miRNAs discovered by RNAseq of 16 different samples taken from Nile tilapia adult tissues and developmental stages. In the first step of this analysis, we clustered miRNAs by their expression profiles using Mapman and identified 40 and 12 clusters considering only expression data from, respectively, adult tissues and developmental stages. Then we predicted putative miRNA-mRNAs interactions with TargetScan (version 6) and applied filters based on the evolutionary conservation of interactions and on the experimental evidence of expression of the target mRNAs (SRP009911) to create a collection of high confidence 11,168 miRNA-mRNAs interactions among 326 known miRNAs, 60 novel miRNAs and 3,805 mRNAs. Next, we performed a functional enrichment analysis for each set of targets of each miRNA within each of the 52 clusters using g:Profiler with Gene Ontology (GO) terms and, taking into consideration shared targets or similar enriched GO terms, we could infer putative functions to the novel miRNAs such as, for example miR-n483-5p predicted with MYOG gene, which can be involved in muscle development processes ("skeletal muscle tissue development" and "muscle organ development" terms). For miR-n941-5p, we identified a putative regulatory role in muscle dysfunctions ("Abnormal muscle tone" and "Hypopigmentation of the skin" terms). Finally, we can infer the miR-n741-5p regulatory roles at liver ("liver development, lipid metabolic process and fatty acid biosynthetic process" terms).

This work was supported by financial grant from São Paulo Research Foundation (FAPESP).

Tityus serrulatus venom gland: new sodium channel toxins through RNA-Seq

Ana Paula Vimieiro Martins¹, Flávia de Faria Siqueira², Evangelides Kalapothakis¹

¹ Federal University of Minas Gerais, ² Federal Institute of Minas Gerais

According to Brazilian Ministry of Health, *Tityus serrulatus* (Brazilian yellow scorpion) is the species responsible for most of severe accidents in Brazil, which can lead to death. For more than one decade, the number of scorpion accidents in Brazil is increasing and for the moderate and most severe cases, the antivenom is the only effective treatment available. However, its production is very expensive and most of its antibodies are produced against venom compounds that are not toxic to mammals. There are many researches trying to solve or reduce these problems, and all of them request deep scorpion venom knowledge. High throughput transcriptomic sequencing technologies (RNA-Seq) have enabled great progress toward this. Although *T. serrulatus* venom has been extensively studied, many of its elements have not been described or characterized. RNA-Seq technology have already been performed for others scorpion species, but this is the first time it is applied for the transcriptome of Brazilian yellow scorpion. It is evident the necessity of deeper studies about *T. serrulatus* venom toxins, especially due to its medical relevance and as an important source of biotechnological tools. At this work, it was done a partial transcriptome of *T. serrulatus* telson. A unique MiSeq run was performed for cDNA sequencing and downstream analysis proved that variable assembly parameters affect the diversity of identified transcripts. The transcriptome herein described was efficient to identify 13 possible new NaTx with primary structure highly different from those of neurotoxins previously described for the species. Sodium channel toxins (NaTx) are known by their important role in lethality due to scorpionism. These molecules are potential target for immunization studies. This result agrees with the diversity of NaTx reported by recent proteomic researches. Based on this transcriptome, many other venom molecules are going to be annotated and added to previous biochemical information to extend the knowledge about *T. serrulatus* venom.

Funding support: FAPEMIG, CNPq, CAPES

PiRNA signatures of adjacent to tumor tissue as potential biomarkers of gastric carcinogenesis

André M Ribeiro-dos-Santos, Tatiana V Sandoval, Pablo Pinto, Amanda Vidal, Arthur Ribeiro-dos-Santos, Paulo Assumpção, Mônica Assumpção, Sâmia Demachki, Sidney Santos, Ândrea Ribeiro-dos-Santos, Sandro J de Souza, Fabiano Moreira

Instituto de Ciências Biológicas, Universidade Federal do Pará, Belém, PA, 66075-110, Brazil; Hospital Universitário João de Barros Barreto, Universidade Federal do Pará, Belém, PA, Brazil; Núcleo de Pesquisa em Oncologia, Universidade Federal do Pará, Belém, PA, Brazil; Brain Institute, Universidade Federal do Rio Grande do Norte, Natal, RN, Brazil

According to cancer field effects concepts, tissues adjacent to tumors carry molecular alterations, nevertheless normal appearance. These changes have been described both in genetic and epigenetic levels for different tumors, including gastric cancer. Regardless of these alterations, adjacent to tumor samples are still been used as non-cancer control when aiming to discover new biomarkers. PiRNAs are small non-coding RNA (25-33nt) that interacts with PIWI protein, silences transposable elements maintaining genome stability during early phases of development and also may act on post-transcription gene regulation. We investigated the piRNA expression profile in paired tumor and adjacent to tumor samples and in gastric sample from non-tumor patients to identify piRNA as cancer biomarkers. A total of 24 samples were analyzed, including eight tumor (GC) and eight adjacent to tumor (AT) paired samples and eight non-cancer gastric tissue (NC). The piRNA was sequenced using Illumina MiSeq platform and statistical analysis was conducted in R using DESeq2 package to identify differentially expressed transcripts ($q\text{-value} \leq 0.05$ and $|fold-change| > 3$). Assuming adjacent to tumor tissues potentially carry molecular alterations, we compared the expression profile among NC and GC to identify biomarkers capable to differentiate non-cancer to cancer tissue. This analysis identified 12 piRNA differentially expressed, being two under expressed and ten over expressed in GC. In order to evaluate the piRNA expression differences between non-cancer, cancer and adjacent to cancer tissues, the expression profile of AT samples was compared to both NC and GC samples. Twelve piRNA were found differentially expressed when compared to GC and eight to NC. Among the differentially expressed piRNAs, three were found under expressed both in AT and NC samples when compared to GC. In a standard biomarker discovery analysis that compares just AT to GC samples, only these three markers would be able to differentiate a non-cancer tissue from cancer. The remaining nine markers would only differentiate adjacent to tumor tissue from cancer. The standard analysis would also lose nine other potential markers that differentiate non-cancer tissue to cancer. The results were similar to previous findings in miRNA and methylation studies and further support the gastric cancer field effect theory. The consequences of using adjacent to tumor as normal control in molecular analysis involves not only missing biomarkers as well as lose the capability to truly differentiate gastric non cancer from cancer tissues. Thus, the standard strategy to identify molecular biomarkers of cancer should be revised.

Analysis of the lincRNA transcriptome in the accessory olfactory system

Antônio Pedro C. B. R. Camargo¹, Marcelo F. Carazzolle¹, Fabio Papes¹

(1) *Genomics and Expression Laboratory, Institute of Biology, University of Campinas,
Brazil*

Financial support provided by FAPESP

The olfactory system is a sensory system capable of detecting environmental chemical cues, leading to the sensation of an odor and/or behavioral and endocrine changes. In order to perform these functions, this system comprises two olfactory organs in mammals, the main olfactory epithelium (MOE) and the vomeronasal organ (VNO), found in the nasal cavity. The VNO is responsible for detecting intra and inter-species stimuli and for initiating innate behaviour, such as sexual, aggressive and social. Recently, a huge variety of long non-coding RNA (lncRNAs) has been discovered in several tissues, playing roles in the regulation of gene expression and development. Given the unique properties of the process by which genes coding for vomeronasal receptors are expressed in the VNO sensory neurons, we hypothesize that lncRNAs might be involved in such regulation. In order to unveil intergenic long non-coding RNAs (lncRNAs) that could be participating in the process of VNO neurons differentiation, we developed a bioinformatics pipeline to identify and functionally annotate lncRNAs preferentially expressed in this organ. Using public RNA-Seq libraries from eight tissues, including the VNO and MOE, we constructed a transcriptome atlas of mice using differential gene and transcript expression analysis based on the mouse Ensembl reference genome that is being utilized for searching lncRNAs. Bioinformatics tools are being used for predicting the coding potential of a transcript using information about the nucleotide composition, evolutionary pattern, ORF length and similarity against known proteins and protein domains. Non-coding transcripts that are differentially expressed in the VNO will be selected to further inspection, both *in silico* and *in vitro*. We will predict the secondary structure of these lncRNAs as well as their possible interactions with proteins and other RNA molecules in order to try to infer their functional role in the tissue. Wet lab experiments, such as real-time PCR and *in situ* hybridization, will provide more information concerning the levels of expression and the spatial localization of the selected transcripts in the VNO. We expect to discover lncRNAs that participate in VNO neuron differentiation, contributing to their unique gene expression and physiological properties, ultimately resulting in the generation of innate behaviours in mice.

Identifying gene clusters in the genome of *Trypanosoma cruzi*

Prado, W. S.; Reis, A. L. M.; Castro, T. B. R.; Franco, G. R.;

Universidade Federal de Minas Gerais

The World Health Organization still considers American Trypanosomiasis, Chagas Disease, as a neglected tropical disease. The causative agent, *Trypanosoma cruzi*, has a complex digenetic life cycle and diverges significantly from other eukaryotes regarding transcriptional regulation and genomic organization. Similar to prokaryotes, transcription is polycistronic in this parasite; however, translation depends on monocistronic mRNAs, which are processed individually from the polycistronic transcript by coupling two reactions: the spliced leader trans-splicing at the 5'-end and polyadenylation at the 3'-end. Only a few promoters have been described for RNA Polymerase II and the boundaries of polycistrons are yet to be defined in this parasite. Some studies in other organisms use the Pearson's correlation coefficient to verify expression patterns and identify networks of genes spatially close in the genome. Nevertheless, such approach has not yet been applied to *T. cruzi*. Based on the assumption that most mRNAs from a common polycistronic transcript should have a related expression, the main goal of this study is to identify gene clusters, as an attempt to rebuild polycistrons, even partially. Thus, we developed a Python script that integrates the genome annotation to RNA-seq data originated from epimastigotes of the CL Brener strain exposed and not exposed to 500 Gy of gamma radiation. In the experimental group, total RNA was extracted 4, 24 and 96 hours after exposure to ionizing radiation and all samples had two biological replicates. The program uses a sliding window without a fixed size to compare the expression of adjacent genes in the chromosome. If the genes are correlated, then they are assigned to the same cluster. Using this approach, we found 1,859 gene clusters in the Esmeraldo-like haplotype of the CL Brener strain of *T. cruzi*, which covers 5,448 out of the 10,597 genes annotated. Most clusters are dicistrons (58, 7%) or tricistrons (21%). The maximum size of a cluster was 19 genes. Several clusters in the range of 5 to 10 genes consisted of surface proteins (mucins, trans-sialidases and MASP), which are involved in the evasion of the host immune system by the parasite. If virtually all protein-coding genes are polycistronically transcribed, finding 51.4% of them in clusters seems to be an underrepresentation. Therefore, these findings support the hypothesis that regulation of gene expression in *T. cruzi* occurs post-transcriptionally.

Financial support: CNPq, CAPES, FAPEMIG, UFMG

Molecular diversity of the venom gland from Peruvian scorpion *Hadrurooides* *lunatus* revealed by transcriptome analysis.

Thiago Mafra Batista¹, Clara Guerra Duarte³, Anderson Oliveira do Carmo²,
Benigno Tintaya⁴, César Bonilla⁴, Evanguedes Kalapothakis², Glória Regina
Franco¹, Carlos Chavez-Olortegui¹

Departamento de Bioquímica e Imunologia, Universidade Federal de Minas Gerais¹,
Departamento de Biologia Geral, Universidade Federal de Minas Gerais², Fundação
Ezequiel Dias³, Instituto Nacional de Salud, Lima, Peru⁴

The venom of Peruvian scorpion *Hadrurooides lunatus*, which is the most medically relevant species in Peru have been poorly characterized. These venoms can hold diverse undiscovered bioactive peptides, enzymes and toxins. The identification of new toxins and peptides have pharmacological, biotechnological and medical implications. In this work, we describe the analysis of the transcriptome from venom glands extracted from the Peruvian scorpion *Hadrurooides lunatus*, performed with the Illumina MiSeq platform. Using a high-throughput methodology, we describe the universe of transcripts and putative peptides that are expressed in this venom glands. The library was prepared using the TruSeq RNA Library Kit, with enrichment for polyadenylated transcripts and selection for fragments with mean size of 200 bp. Sequencing was performed using the Illumina MiSeq V3 Kit, generating paired-end reads of 301 bp. Twenty eight millions of reads were produced and collapsed using PEAR v0.9.8, forming single reads, and transcripts assembly was carried out using rnaSPADES v3.9.1. In total, 231,685 transcripts were assembled and translated into 29,699 putative peptides using TransDecoder. The search for sequence similarity using Blastp against a database composed by 6534 toxins and venoms deposited at Uniprot revealed 668 peptides with similarity with at least one protein from database. We identified 400 possible proteases and more than 300 molecules, including toxins and peptides with high biotechnological and medical values. Non buthidae scorpions are known to contain a large amount of bioactive peptides. In this work, we describe some of the identified enzymes, antimicrobial peptides, toxins for potassium, sodium and calcium channels with a high degree of identity (over 50%) with molecules from other scorpions and arthropods. The results obtained in this work represent the first landscape of components of a scorpion venom belonging to the Iuridae family, revealing the complexity of molecules expressed in this tissue, with great potential for future uses in medical and evolutionary studies.

Trypanosoma cruzi coding transcriptome in response to gamma radiation

Pereira MA¹, Imada EL¹, Grynberg P², Vieira HGS³, Kaczorowski D³, Macedo AM¹, Machado CR¹, Franco GR¹

¹Departamento de Bioquímica e Imunologia, UFMG, Belo Horizonte, Brasil; ²Embrapa Recursos Genéticos e Biotecnologia, Brasília, Brasil; ³Garvan Institute of Medical Research, Sydney, Australia

Trypanosoma cruzi, the etiologic agent of Chagas disease, is a kinetoplastid organism highly resistant to DNA damage caused by ionizing radiation. After a dose of 500 Gy of gamma rays, the genomic DNA is fragmented. Interestingly, the parasite is able to restore the chromosomal bands pattern in less than 48 hours. Previous studies using microarrays and 2D PAGE followed by MS/MS analyzed how gamma rays affect *T. cruzi* gene expression. Microarray analysis showed that transcripts related to basal metabolic functions were down-regulated. In contrast, the up-regulated category was mainly composed by obsolete sequences, hypothetical proteins and Retrotransposon Hot Spot genes. Proteomic analysis indicated that active translation is essential for the parasites recovery from ionizing radiation damage. The presence of shorter protein isoforms after irradiation suggests the occurrence of post-translational modifications and/or processing in response to gamma radiation stress. Our study aims to analyze the gamma radiation effect on the *T. cruzi* transcriptional profile by high-throughput RNA sequencing (RNA-Seq) and to increase our knowledge on the molecular mechanisms related to the parasite resistance do ionizing radiation. Epimastigote cells from CL Brener strain were exposed to a dose of 500 Gy in a cobalt (60Co) irradiator. Total RNA was extracted from non-irradiated (control sample) and irradiated cells (4, 24 and 96 hours post-irradiation). Two biological replicates were produced for each condition. RNA-seq paired-end strand specific libraries were prepared using poly(A) enrichment/dUTP incorporation and sequenced on the Illumina HiSeq 2500 platform. Approximately 210 millions paired-end reads were obtained. FastQC was used for quality control. Reads were edited *in silico* to remove ERCC92 sequences, adapters and low quality regions. In order to generate a reference transcriptome all samples including biological replicates were combined into a single RNA-Seq data set and assembled by Trinity with different k-mers (25, 27, 29 and 31-mers). Trinity 31-mers assembly was selected to downstream analysis after evaluation by Transrate. A total of 75,798 transcripts (44,773 genes) were generated with an average contig length of 717.52 bp and N50 value equal to 1584 bp. Next steps include transcripts redundancy decrease, transcriptome annotation, differential expression and functional analysis. This study will help to understand how the parasite can handle such a harmful stress.

Financial Support: Capes, CNPq, FAPEMIG, Garvan Institute

Combined genome guided and long reads assembly of the *Coffea arabica* transcriptome

Rezende P. M., Ribeiro T. H. C., Fernandes-Brum C. N., Schumacher P. V., Ferrara-Barbosa B. C., Chalfun-Junior A.

Laboratory of Plant Molecular Physiology, Federal University of Lavras

Coffee represents a great source of income for several countries and Brazil poses itself as the world's biggest producer and exporter of this commodity. The genus *Coffea* has more than 124 species, however, only two are economically relevant: *Coffea arabica* and *Coffea canephora*. *Coffea arabica*, the only tetraploid species, originated from a crossing event between *Coffea canephora* and *Coffea eugenioides*. The elevation of global temperatures caused by climate change is a threat to coffee production and is estimated that it may cause losses of up to \$2,9 bi by 2020. Given this scenario, it becomes necessary to take measures to assure the global production of coffee. The study of the coffee genome and transcriptome can provide the basis for the improvement of coffee production and quality, through the development of new cultivars using techniques such as genetic engineering. However, only the genome of *C. canephora* has been sequenced so far. Thus, in this study, we used paired-end RNAseq libraries of two *C. arabica* cultivars ("Acauã" and "Catuaí Vermelho"), grown under to two different temperature ranges (19/23 °C, 26/30 °C), along with the *C. canephora* genome and EST sequences, from the CAFEST database, to reassemble the *C. arabica* transcriptome. The RNAseq libraries were aligned to the *C. canephora* genome, using the aligner STAR v2.4.2, and approximately 85% of coverage with this genome was obtained. The transcriptome assembly was performed with Trinity v2.2.0 tool, using as a reference genome the libraries assembly, and to enrich the assembly we also added as long reads ESTs from a public database of coffee ESTs (CAFEST). As results, 108940 putative genes and 144480 putative transcripts were identified. The N50 length, the statistics that define assembly quality, was 1781bp for the transcripts and 1296bp for the biggest transcripts of each gene. Hence, from these results we demonstrate that the strategy used in this study for the transcriptome assembling was wide and robust, and it can be used as a genomic resource for future investigation on *C. arabica*.

Assembly, identification and characterisation of sugarcane transcripts

Rezende P. M., Ribeiro T. H. C., Schumacher P. V., Lima A. A., Chalfun-Junior A.

Laboratory of Plant Molecular Physiology, Federal University of Lavras

Sugarcane (*Saccharum officinarum L.*) stands out as an important crop due to its role in sugar and ethanol production, which are widely consumed worldwide. Brazil is the world leader in sugar production from sugarcane, being responsible for 38.7% (1.9 billion tons) of the world production in 2014. Unlike other important crop species, such as Maize (*Zea mays L.*), Sorghum (*Sorghum bicolor L. Moench*), and Rice (*Oryza sativa*), sugarcane haven't had its genome sequenced so far, what limits the understanding of important biological processes from the molecular point of view. Among these important processes is flowering, which should be avoided in sugarcane since it requires a great deal of energy, reducing the sugar content of the sugarcane stalks. Thus, a better understanding of the molecular mechanisms regulating sugarcane flowering induction is crucial for the reduction of flowering intensity in sugarcane and the development new cultivars less susceptible to flowering. This study aimed to perform an assembly strategy, the identification, and the characterisation of sugarcane transcripts available in public databases. Transcript assembling was carried out using the software Trinity, using the SUCEST (Sugarcane Expressed Sequence Tag) database and sugarcane reads generated from a high-throughput sequencing platform as input data. After the identification of the reads, the coding regions and the protein sequences of the candidates genes were predicted using the software TransDecoder. The predicted proteins were aligned against the protein database SWISSPROT. The alignments, a conserved domain analysis, and the construction of phylogenetic trees using flowering gene sequences from other species, allowed the identification of two important sugarcane flowering genes, FD and FT (*Flowering Locus T*). 151389 genes and 170756 transcripts were assembled, and the N50 value for these transcripts was 929. The alignment against the SWISSPROT database showed that 137126 transcripts displayed an e-value below 10, what corresponds to 97 % of the transcripts identified. From these 137126 transcripts, 57% showed an alignment coverage higher than 80%. The conserved domain analysis indicated the presence of 14 sugarcane sequences that possess the conserved bZIP (*basic Leucine Zipper*), present in the FD gene from other species, and one of these sequences was shown to be a putative sugarcane FD based on the phylogenetic analysis. On the other hand, for the FT gene, 11 sugarcane sequences were found to possess the PEBP (*Phosphatidyl Ethanolamine-Binding Protein*) domain, which characterises FT genes, and the phylogenetic analysis showed six of them were putative FT genes, although four of the sequences were found to be partial sequences. Thus, these results show that the approach used in this study can be used to identify putative genes related to important biological process of sugarcane.

Caprin-1 binding profile to target RNAs via enhanced CLIPseq

Ciamponi F.E.¹, Migita N.A.¹, Van Nostrand E.², Lovci M.T.¹, Alonso L.¹, Aigner S.², Yeo G.W.², Massirer K.B.¹

1 - Center for Molecular Biology and Genetic Engineering, University of Campinas, Campinas, Brazil; 2 - Dept of Cellular and Molecular Medicine University of California San Diego, La Jolla, USA

Caprin-1 is a cytoplasmic RNA binding protein (RBP) expressed in most mammalian tissues and is particularly abundant in neurons as part of RNA granules. This RBP binds to RNA targets in mRNPs complexes and contains a conserved RG-rich domain, which is related to cellular granules formation and aggregation of proteins. Caprin-1 is also a component of high-density cytoplasmic granules called stress granules (SG), a structure comprised of RNA and protein in a tightly packed aggregate, which has been associated with regulation of protein expression via stalling of translation pre-initiation complexes in target RNAs. Several neurodegenerative diseases have been associated with formation of stress granules. While toxicity of granules is still under debate, we wanted to better understand Caprin-1 function in SGs and its implications in the cell. Since ectopic expression (EE) of CAPRIN1 in cells is sufficient to induce the formation of SGs, we performed enhanced crosslink immunoprecipitation sequencing (eCLIPseq) on human HEK293T cells subjected to CAPRIN1 EE. This approach allowed us to define the target RNAs bound to the protein of interest and to profile potential binding sites to the target RNA (binding sites). We complemented this approach with other high-throughput techniques such as RNA-immuprecipitation sequencing (RIPseq) and standard RNA sequencing (RNAseq) to allow a biologically relevant set of target genes to be used in further wet lab experiments. Our results revealed that Caprin-1 binds to 1720 different RNA targets, comprised mostly of mRNAs, in multiple locations, indicating a "coating" behavior in the binding activity, with a bias towards the CDS and 3'UTR regions. Based on over-representation of sequences capable of forming G-quadruplex structures in binding sites, we propose that Caprin-1 binds to RNA via this secondary RNA structural motif. We also observed that 174 binding targets were abundant in the RIPseq, gene ontology profiling of this subset showed enrichment in nucleotide binding and RNA-metabolism related classes. Complementary analysis from RNAseq also indicates that binding targets of Caprin-1 are mostly up-regulated after stress granule induction. In summary our findings suggest that Caprin-1 binds to a structural RNA motif, which can lead to stabilization of a set of mRNA transcripts enriched in RNA-related targets and regulate their expression.

Financial Support: FAPESP, CAPES, FAEPEX-Unicamp, NIH

Comparison of the Expression profile between embryogenic and non-emбриogenic *Coffea arabica* L. calli through RNA-Seq data analyses using combination of DE algorithms

Thales Henrique Cherubino Ribeiro¹, Wesley Pires Flausino Máximo², Kalynka Gabriella do Livramento², Anderson Tadeu Silva², Antônio Chalfun Júnior¹, Luciano Vilela Paiva²

1 *Laboratory of Plant Molecular Physiology, Federal University of Lavras*

2 *Central Laboratory of Molecular Biology, Federal University of Lavras*

Coffea arabica L. is the main source of one of the most important beverages worldwide; more than 2 billion cups of coffee are consumed every day. Coffee trees are grown on more than 10 million hectares of tropical land, and Brazil is the world leader producer and exporter. Traditionally, coffee seedlings are produced from seeds, and undesirable traits can arise during the process of seed formation due to genetic recombination. On the other hand, the use of seedlings regenerated from somatic embryogenesis (SE) can be a more effective method for growing coffee once every new plant generated will be essentially a clone of the mother plant. Thus, in this work, we sequenced and compared the transcriptomic profiles of embryogenic and non-emбриogenic calli from *Coffea arabica*. After a quality control, approximately 92% of the remaining 59,405,225 Illumina single-end reads were successfully mapped on the *Coffea canephora* genome using STAR aligner version 2.4.2a. Then, Cufflinks package was used to assemble the transcriptome and to identify new putative genes and isoforms. Differential expression analyses were carried out using a combination of the programs cuffdiff, edgeR and DESeq2. Those genes that showed at least a 2-fold expression change between the conditions, false discovery rate (FDR) below 5% of significance and meet the previous two conditions in at least two of the three programs were considered differentially expressed. We found 3,882 novel loci in the assembled transcriptome and 6,986 differentially expressed genes (DEG) between the two calli types, being 2,170 of those genes transcriptionally more active in embryogenic calli. The investigation and experimental validation of the role of those DEG will be the foundation for future research on the molecular control of SE in *Coffea arabica*, guiding us toward the understanding of the mechanisms involved in the transition between non-emбриogenic calli to embryogenic ones.

Financial support: CAPES and CNPq

Genome-wide identification and *in silico* characterization of microRNAs and their targets in *Ananas comosus* L.

Thales Henrique Cherubino Ribeiro¹, Pâmela Marinho Rezende¹, Laurence Rodrigues do Amaral², Matheus de Souza Gomes² and Antonio Chalfun Júnior¹

1 Laboratory of Plant Molecular Physiology, Federal University of Lavras

2 Laboratory of Bioinformatics and Molecular Analysis (LBAM), Federal University of Uberlândia - Campus Patos de Minas (UFU)

Pineapple (*Ananas comosus* (L.) Merr.) is an economically important tropical fruit and the most important crop in the Bromeliaceae family. Due to pineapple's close proximity to crops of the Poaceae family such as rice, corn, sugar-cane and wheat, the investigation of regulatory and developmental elements in this species is useful for crops enhancement programs. One of the key regulatory elements are microRNAs (miRNAs). They are small non coding RNAs (approximately 21 to 24 nucleotides in length) that negatively regulates gene expression by guiding a RNA-induced silencing complex (RISC) to a complementary target messenger RNA (mRNA) and directly promote post-transcriptional gene silencing. In this work we identified and characterized the mature miRNAs, precursor miRNAs and target genes in the genome and transcriptome of *A. comosus* using a series of Bioinformatics steps. Firstly, we applied a robust and adapted algorithm described by de Souza Gomes et al., (2011) to the genome of *A. comosus*. Following steps including homology searches, secondary structure prediction, and minimal free energy evaluation, the algorithm identified 277 putative mature miRNAs and their respective precursors. Then, all of those sequences were manually evaluated considering their homology results to known and conserved miRNA sequences, thermodynamic and structure characteristics of pre-miRNAs, and phylogenetic analyses. After the manual curation the remaining 102 miRNAs and their precursors were classified. Then we performed de novo assembly of the transcriptome from single-ended reads of pineapple's apical meristem and roots using the software Trinity. The web-tool psRNATarget, available at <http://plantgrn.noble.org/>, was used to identify miRNA's target sequences on these transcriptomes. The 102 identified miRNAs in pineapple had 575 target transcripts in meristem samples and 582 targets in roots. A better characterization of the miRNAs and their target genes will help the understanding of how tissue-specific gene regulation controls the pineapple's development and physiology.

Financial support: CAPES and CNPq

Lsm-bound antisense RNAs play role in *Halobacterium salinarum* NRC-1 transposition regulation

Alan P. R. Lorenzetti¹, José Vicente Gomes-Filho¹, Livia S. Zaramela², Felipe ten Caten¹, Ricardo Z. N. Vencio³, Tie Koide¹

¹Departamento de Bioquímica e Imunologia, Faculdade de Medicina de Ribeirão Preto, USP, Ribeirão Preto, Brasil, ²Division of Host-Microbe Systems & Therapeutics, University of California, San Diego, USA and ³Departamento de Computação e Matemática, Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto, USP, Ribeirão Preto, Brasil

Insertion sequences (IS) are widely spread in bacteria and archaea and are well known for their impact in host genomes. IS transposition events may inflict gene disruptions, massive deletions and even the creation of new features by rearranging stretches of DNA. Therefore, transposition are usually kept at low rates by several mechanisms, such as the translational repression of transposase-encoding mRNAs. Hfq binding proteins associated to antisense RNAs (asRNAs) were reported to inhibit the translation of transposase-encoding transcripts from IS200 family in *Salmonella*. The repression occurs by Hfq-bound asRNA base pairing on the 5' UTR and transposase (tnpA) first codons, preventing the 30S ribosomal subunit from attaching to Shine-Dalgarno sequence. Lsm, the Hfq homologue in archaea, is an important agent in post-transcriptional regulation of protein-coding mRNAs in *Halobacterium salinarum* NRC-1. Nevertheless, interaction between Lsm-bound asRNAs and transposase-encoding transcripts is still unexplored in this haloarchaeon. Starting from RNA immunoprecipitation sequencing (RIP-seq) data, we aligned reads to reference genome and generated read counts per position for pulled-down Lsm and control libraries. We also computed fold change per position based on read counts, in order to avoid overestimation of interaction regions. This allowed us to identify Lsm-bound RNAs interacting within IS transcripts in *H. salinarum* NRC-1 grown at reference condition. We found antisense interaction regions at the 5' end of IS200/IS605 family transcripts, frequently spanning the beginning of transposase-encoding sequences, suggesting a regulation mechanism similar to that described in bacteria. These preliminary results encourage a genome-wide analysis to evaluate whether Lsm-bound asRNAs are bona-fide IS transposition regulators in archaea, by the means of comparing the expression levels of transposase genes in Lsm knockout mutants and different growth conditions.

Funding support: CAPES and FAPESP.

Non coding RNAs in *Coffea canephora* genome: Identification by similarity.

Samara M. C. de Lemos¹, Alexandre R. Paschoal¹, Douglas S. Domingues^{1,2}

¹Graduation Program in Bioinformatics – PPGBIOINFO – Universidade Tecnológica Federal do Paraná, Cornélio Procópio; ²Department of Botany, Instituto de Biociências, Universidade Estadual Paulista, Rio Claro

Non-coding RNAs (ncRNAs) are an important component of genomes and transcriptomes of eukaryotes. They are preferentially generated in intronic and intergenic regions of the genome. ncRNAs can comprise microRNAs (miRNAs), transporter RNAs (tRNAs), small nucleolar RNAs (snoRNAs) and small nuclear RNAs (snRNAs). Fundamental processes for plant physiology and development, such as flowering and fruit ripening are regulated by ncRNAs, which motivates the identification of ncRNAs in plant genomes. Coffea genus has at least a hundred species and some of them has an important role in agricultural industry. Brazil is the leading producer and second largest consumer market of coffee. Recently, the genome of Robusta coffee (*Coffea canephora*) became available for analysis. This species accounts for one third of the world production. This study aimed the identification of ncRNAs in the newly sequenced *C. canephora* genome. Intergenic and intronic regions were extracted using a pipeline developed with *Shell* and *Perl* scripts. We BLASTed these regions against ncRNA public data from 31 plant species available in ENSEMBL Plants (version 32) and applied a *Perl* script to filter hits above 80% of identity and coverage. Using these criteria, we identified 47 hits in *C. canephora* introns, 13 miRNAs, 25 tRNAs, 7 snoRNAs and 2 snRNAs. From intergenic regions we obtained 525 alignments. They were classified as 273 snRNAs, 231 rRNAs, 14 ncRNAs and 7 miRNAs. The 3 species that presented most hits in coffee were *O. sativa* (rice), *S. tuberosum* (potato) and *S. lycopersicum* (tomato). These preliminary results provide a starting point for the characterization of a genomic component that is a protagonist in several agronomical traits.

Acknowledgement: Arabica Coffee Genome Consortium.

Distinguishing coding and non-coding RNA sequences and improving its functional annotation using machine learning approaches

Thaís de Almeida Ratis Ramos¹, Daniel Miranda de Brito², Raul Arias-Carrasco³, Leonardo Vidal Batista², Thaís Gaudencio do Rêgo², Vinicius Maracaja-Coutinho^{3,4,5}

¹*Universidade Federal do Rio Grande do Norte*, ²*Universidade Federal da Paraíba*,

³*Universidad Mayor*, ⁴*Instituto Vandique*, ⁵*Beagle Bioinformatics*

Non-coding RNAs (ncRNAs) are important players in the cellular regulation in organisms from all domains of life. Its investigation is already routine in every transcriptome or genome project. Two key steps on the predicting process and functional assignation of ncRNAs, are (i) the ability to distinguish coding and non-coding sequences, followed by (ii) a functional assignation of RNA families based on sequences similarity searches or secondary structure predictions. Here, we applied different machine learning approaches in order to distinguish coding and ncRNA sequences, and to functionally predicted ncRNAs into known RNA families. The coding potential prediction was developed using different randomly selected sets of ncRNA sequences, extracted from Rfam database; and human RefSeq protein coding genes. Coding and ncRNAs had their tri-nucleotides counts analyzed using three different equally divided sets of 200, 400 and 1000 instances. For the functional classification of ncRNAs, we performed multiple alignments using sets of (i) 100 sequences and secondary structure models (SSMs) in Dot-Bracket Notation from 10 different families; (ii) 200 sequences and SSMs from 20 families; and (iii) three different sets of 500, 1000 and 2000 ncRNA sequences from 50 families. All RNA families were randomly selected from Rfam. Sequences and SSMs were filtered using a maximum similarity cutoff between them of 80% (Levenshtein distance); and a maximum length of 400nt. Then, we analyzed the counts of mono-, di- and tri-nucleotides on the primary sequences. Next, multiple alignments were performed using Clustal Omega and MARNA, respectively for primary sequences and SSMs. Finally, different classification tests were performed, using Naive Bayes, SMO, IBK, Multilayer Perceptron and Random Forest through WEKA tool. The coding potential evaluation using 200, 400 and 1000 sequences presented accuracies reaching 99%, 99% and 99.2%, respectively. The functional assignation of ncRNAs using 10 and 20 families, revealed results with an accuracy reaching 99% and 98.5%, respectively. Tests performed using 50 ncRNA families, resulted in an accuracy of up to 94.2%. These results outperforms predictions available in literature, which used a maximum of 25 RNA families. Tests were also performed in order to predict new ncRNAs families in different transcriptome data, opening new opportunities for the development of novel tools for nucleotides coding potential prediction and for the functional classification of ncRNA sequences. Future directions consists on the evaluation of our methodology performance using different sets of specie-specific nucleotide sequences and SSMs.

Proposal of a data mining pipeline to improve bacterial small RNA prediction

Fabio Reinoso Vilca, Sabrina de Azevedo Silveira, Fabio Ribeiro Cerqueira

Departamento de Informática, Universidade Federal de Viçosa, Viçosa-MG

In the last years, the discovering of novel bacterial small RNAs (sRNAs) became relevant due to their essential roles in important cellular activities. Such molecules are key for a number of mechanisms such as: Regulation of outer membrane protein expression, iron homeostasis, quorum sensing, and bacterial virulence. Prediction of sRNA is a challenging issue in bioinformatics, i.e., the current computational tools have low precision and sensitivity. However, the developments of predictive methods are of fundamental importance to narrow the number of costly and time-consuming sequence validations on the laboratory workbench. In this work, we extract different kind of features of a putative sRNA sequence to perform the prediction task. Some important features are: Sequence-based features, secondary structure features, base-pair features, triplet sequence structure, and structural robustness. Additionally, we apply the InformationGain feature selection algorithm to select the best 25 features over the initial 251 set of features. Our preliminary results show that the most relevant features are those related to the secondary structure. Our dataset is composed of 794 known and experimental-validated sRNAs obtained from the BSRD database, as well as 1219 different kind of non-sRNA sequences such as: Shuffled sequences generated from real sRNAs preserving the di-nucleotides composition, and other noncoding types of RNAs (tRNAs, rRNAs) from 5 different bacteria substrains pertaining to 4 different families, including one cyanobacteria. Finally, we apply the random forest learning algorithm for the classification task. The current results of our approach reached an accuracy of 80.63%, an specificity of 82.2%, and an sensitivity of 78.2%.

Differential Gene Expression Analysis of Placentas from *Mus musculus* Exposed to Different Stress Conditions

André Barbosa, Ana Carolina Tahira, Helena Brentani

*Inter-institutional Post-Graduation Program on Bioinformatics of University of São Paulo,
Institute of Psychiatry of University of São Paulo*

Human and animal studies have shown that maternal stress during pregnancy is associated with development of different diseases in offspring. Stress conditions as diabetes, obesity, anxiety have been associated with increased risk for different neurodevelopmental disorders, such as Autism Spectrum Disorders (ASD). Placenta is the organ that accomplishes maternal-fetal mediation in response to adverse conditions, being responsible for providing intrauterine homeostasis which is essential for normal fetal development. Thus, regulatory process undertaken by several genes in placenta seems to be important for the development of the fetal brain. The aim of this study was to evaluate if two different stressful conditions, hypoxia and high-fat diet (HFD), affect the expression of the same genes in mouse placentas, and which biological processes these genes are involved. Transcriptome datasets of hypoxia and high-fat diet were obtained from Gene Expression Omnibus (GEO) database and pre-processed for correction of background and normalization. The differential expression analysis was performed by comparing Case vs Control using Significance Analysis of Microarray (SAM) implemented in R. Finally WebGestalt tool was used to access which biological processes the genes differentially expressed were enriched. Analysis showed differential expression of 58 genes in HFD condition and 861 genes in hypoxic condition. Interestingly there was no overlap of any differentially expressed gene between the two conditions. HFD enrichment analysis showed that these genes are present in pathways of regulation, transport and uptake of neurotransmitters, important for fetal brain development since they influence maturation of neuronal circuitry. Hypoxia condition presented enriched genes for pathways of damage response and DNA replication, as well synapse and dendrites projection. Hypoxia also presented genes enriched for phenotypes conditions such as abnormal development of the nervous system and neurodegeneration.

Supported by CAPES and FAPESP

Transcriptional landscape of *Paracoccidioides brasiliensis*: an isolate presenting no dimorphism shift

Oliveira, L. M.^{1,4}; Desjadins, C.²; Ruiz, J.C.³; Alves, V. S⁴; Baltazar, L. M.⁴; Santos, P. C.⁴; Cuomo, C.²; Cisalpino, P.S.^{1,4}

¹Programa de Pós-graduação em Bioinformática, ICB/UFMG; ²Broad Institute of MIT and Harvard; ³Grupo Informática de Biossistemas e Genômica (Fiocruz Minas); ⁴Laboratório de Biologia de Microrganismos/Laboratório de Biologia Celular de micro-organismos, Departamento de Microbiologia (ICB/UFMG).

Paracoccidioidomycosis, a systemic mycosis of significant medical importance, endemic to Latin America, is caused by thermodimorphic fungi of *Paracoccidioides* species complex. The infection is thought to be contracted by inhalation of fungal propagules and the disease is triggered by the dimorphic shift from conidial to the yeast phase at body temperature. We selected from a clinical isolate of *Paracoccidioides brasiliensis* (Pb339; ATCC32069) under pressure of 400 µg of sulfamethoxazole an isolate that presents no dimorphic shift (YRT, Yeast at Room Temperature). By integrating experimental and computational information the systems biology approach aims at generating new insights to underlying complex dimorphism and virulence mechanisms of *P. brasiliensis*. Here, for the first time we apply advanced next-generation sequencing (Illumina RNASeq) in order to investigate of two temperature-regulated states across the reference isolate Pb339 and the defective isolate YRT. In terms of RNA sequencing we found 5×10^8 reads (PE 2x100, Q Phred ≥ 25), mapping $3,2 \times 10^8$ to a reference genome. Estimates of FPKM values (fragments per kilobase of exon per million aligned fragments) were well-correlated between biological replicates. PbB339 yeast and mycelial phases have distinct transcriptomes presenting 248 and 802 differentially expressed genes, respectively. Oxidative stress pathways were the most enriched. The transcriptome of the defective isolate YRT changes dramatically compared to yeast and mycelia transcriptomes of the reference isolate. The defective YRT at 37°C and PBB339 isolate at yeast phase share a similar transcriptional profile, showing 462 up regulated genes. The oxidative and protein synthesis pathways were over-represented. However, at room temperature YRT transcriptome identified 387 differentially expressed genes when compared with M phase in the same conditions. Oxidative stress pathways were also enriched. Additionally, co-factor binding, ABC transporters and Kinase pathways were found to be over-represented. We uncovered 125 genetic variants (SNP/Indel) in YRT isolate, of which 2 were non-synonymous substitutions and 4 were potentially associated to loss of function. The first mutations for MAPK3 e TAO3/PAG1 genes are involved with mycelial growth regulation in pathogenic fungi. Loss of function related to USP protein (universal stress protein) were found with one hypothetical gene presenting tetratricopeptide repeats and glycosil transferase domains involved with response to environmental and dimorphism phenomena, tRNA2 thiolation and APG9 autophagy function. Here we present the only PbB339 Y and M phase-specific transcriptomes carried out by RNASeq as well as for a dimorphism defective isolate. These analyses will provide new information about thermal dimorphism and morphogenesis currently underexplored in *Paracoccidioides*.

R script to HLA epitope predictor based in matrix frequency: training and performance comparisons

Alessandra Lima da Silva¹, Leandro Martins de Freitas²

¹Institute of Biological Sciences at UFMG and ²Multidisciplinary Institute of Health at UFBA

Epitope prediction assists in the identification of candidate proteins to cause a greater immune response, selecting potential targets for studies and applications in the prevention, treatment, and diagnosis of diseases. The aim of this study was to evaluate the performance of a R script; trained with epitope retrieved from the Immune Epitope Database and Analysis Resource (IEDB) and compared with the NETMHCcons predictor. Peptides with 9 amino acid that binds to the MHC I supertype HLA-A * 02: 01 were selected from IEDB. The search in the IEDB returned 5964 epitopes validated in the database, but only 1022 were established according to criteria. Ligand matrix frequency was prepared using IEDB epitopes and not ligand matrix frequency was prepared using amino acid composition in the UniProt data bank, working as background probability. Using the in-house R script it was possible to select potential targets ligand to the MHC I. 1244 protein sequences of the hybrid strain *T. cruzi* CL Brener were obtained to use as target. Both predictions (R script and NetMHCcons), resulted in the same amount of peptides (633,005). Only peptides with strong binding prediction (cutoff of 0.84 in R script) were selected from the results generated by R script, resulting in 1589 peptides. The comparison with the results generated by NetMHCcons showed that approximately 56% of the peptides showed the same prediction compared with the script. Comparing the strong binding epitopes (0.84 or less) predicted with R script and randomly peptides among 633,005 returned only 2% shared peptides. The above findings are the result of the peptide prediction comparison of a new script in R language with a well-established server, NETMHCcons. The R language script is based on probability matrices, a simple and less sensitive analysis. Even so, 56% of the results were similar to the ones generated by NETMHCcons. The results may be tested further for their effectiveness of stimulating an immune response in both *in vivo* and *in vitro* experiments to support the *in silico* findings.

Patent Mining as a Tool for Innovation Planning and Biodiversity Access: Technologies of Açaí Palm Fruit (*Euterpe oleracea* Mart.).

Heitor Cappato Guerra Silva¹, Letícia de Castro Guimarães², Fabiana Regina

Grandeaux de Melo², Foued Salmen Espindola^{1,2}

¹Institute of Genetics and Biochemistry and ²Intellect Agency, Federal University of

Uberlândia. Uberlândia, Minas Gerais, Brazil.

Research and innovation using natural extracts and their compounds from biodiversity resources, enable breakthroughs in discoveries such of new drugs, foods and cosmetics. Bioinformatics approaches such as text mining of patent databases can reveal a global view of the efforts to access natural products and its bioactive molecules from biodiversity hotspot and thus providing information and opportunities to implement bio-entrepreneurship, public policies and measures for the use and conservation of different biomes and their threat biodiversity. Thus, the aim of this work was search patents of the palm fruit *Euterpe oleracea* Mart. (açaí) using the Thomson Innovation database, that cover world patents, and have many tools to analyze and categorize them. The search strategy was the choice of keywords and setting the search period. Using the keywords *Euterpe oleracea*, açaí and assai, looking for cover all the possibilities and the period established was from 2006 to 2015, considering the publication date. This period contains more than 2338 published patents equating to almost 826 patent families, a patent family is one or more published patent originating from a single original (priority) application. We evaluated the number of published patents by publication year, which suggests an increase of publication in these recent years, top applicants and their profile, Countries that have higher number of published documents. The distribution of documents according with IPC classification, as 61K that refers a hygiene, cosmetics products for human healthy, that have 294 published patent, and the top applicants are all international companies. Besides Text Clustering, that automatically categorize documents through the linguistic analysis of text found in the fields title, abstract and claim, to evaluate the different application areas, and based on same fields we visualize the ThemeScape map, a two-dimensional map displaying the relative relationship of one record to another using common conceptual term, which allows to investigate the technological trajectories. The aim is that the results presented here guides lines and research projects, development and technological innovation related to biotechnology with natural products; seek to improve the idea of protect the work before publication. In addition, using bioinformatics approaches to improve the culture of Innovation at the Brazilian Universities, Institutions and Industry.

Financial Support: CNPq, Fapemig and PROPP/UFU.

mirhunt: an approach to predict microRNA binding sites using different prediction tools

Jaqueline Ramalho^{1,2}, Michelle A. Paz^{1,2}, Iane O. P. Porto^{1,2}, Celso T. Mendes-Junior³,
Erick C. Castelli^{1,2}

(1) Programa de Pós-Graduação em Patologia – Faculdade de Medicina de Botucatu, UNESP, (2) Laboratório de Genética Molecular e Bioinformática, Unidade de Pesquisa Experimental, Bloco 5, Faculdade de Medicina de Botucatu – UNESP and (3) Departamento de Química – Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto, USP

One of the mechanisms for gene expression regulation occurs post transcriptionally by the binding of microRNAs to the mRNA. microRNAs are small noncoding RNA molecules of about 22 nucleotides that usually interacts with the mature mRNA 3' untranslated tail. Their binding occurs through the seed region (usually 2-7 nucleotides) in the microRNA 5' end. This binding might suppress translation or trigger mRNA degradation. The same microRNA may bind to several targets, and one gene might be regulated by several microRNAs, confounding regulatory networks. Several studies have reported the up- and down-regulation of specific microRNAs in pathological contexts. In this matter, it is important to infer which microRNA may influence the expression of clinically relevant genes for future functional studies. The prediction of microRNA targets or binding sites usually requires the use of different prediction tools, mainly because they might predict different targets for a same microRNA/mRNA pair as well as to avoid false-positive interactions. In addition, the use of long mRNA sequences to predict microRNA/mRNA interactions might bias the analyses detecting only the most stable bindings. *mirhunt* is an application designed to automate the searching for human microRNA binding sites by using three different prediction algorithms, miRanda, RNAhybrid and IntaRNA, comparing the results of these methods and applying a scoring system to classify each interaction. The scoring system is based on the strength of each microRNA/mRNA interaction found by comparing it with a database (matrix of 26,414 human mRNA sequences inferred from the annotations provided by the human genome draft version hg38 and 2,588 known microRNAs (miRBase)), which containing the interactions previously detected. Long sequences are managed by fragmenting them on a series of overlapped subsequences and processed independently, maximizing the ability of these algorithms to predict miRNA/mRNA interactions on long mRNA sequences. Reports are generated presenting all the interactions found and the scores for each interaction. The scoring system used minimizes the presence of false-positive interactions, since it considers the results from three prediction algorithms, and also allows the selection of the most stable bindings and most specific bindings by comparing the target results with the *mirhunt* database. For example, *mirhunt* was used to infer the microRNA binding sites to the HLA-G 3'UTR sequence and the most specific bindings detected were related to miR-148a-3p and miR-148b-3p, whose influence on the HLA-G expression profile was functionally proved. The data provided by *mirhunt* may provide background for functional studies. Available: http://www.castelli-lab.net/apps/apps_mirhunt.php.

Financial Support: FAPESP, CNPq
E-mail: castelli@fmb.unesp.br

Alien: A tool for handling sequence alignments

Dhiego Souto, Lucas Bleicher

Universidade Federal de Minas Gerais

Many softwares developed for bioinformatics studies use multiple sequence alignments as an input. To build a multiple alignment is not a trivial task. To generate its global optimum it is required an infeasible time for a computer to calculate in larger cases. So, there are many softwares that build multiple alignments with heuristics and meta-heuristic techniques, which can run in a feasible time although its response is not the optimum. The more accurate is the alignment, the better is the extraction of information from these softwares. Alien implements the following pair alignment algorithms: local alignment algorithm (Smith & Waterman, 1981), also implemented by BLAST; global alignment algorithm (Needleman & Wunsch, 1970). When dealing with pair alignments, is possible to reach optimum in feasible time. Alien also implements a profile hidden Markov model (Eddy, 1994), the Viterbi algorithm (Viterbi, 1967), specific Blosum matrices from any protein multiple sequence alignments and two evaluation methods based on sum of pairs and hmm probabilities. Running the Viterbi algorithm through the profile hmm it is possible to generate a multiple sequence alignment. Alien multiple alignments for short sequences showed good results comparing to Hmmer. Recently ARCA, (Architecture Alignment) was implemented, which is an alignment of proteins architecture. A protein may possess one or more domains, each configuration of these domains is denominated as an architecture of the protein. ARCA searches proteins databases for protein sequences which share the same architecture and align them with the aligners HMMER and MUSCLE. The domains are already aligned by HMMER, and available on Pfam database, but on Pfam these alignments carry only the amino acids related to each domains. So ARCA's job is to gather all the family alignments available on Pfam database, the deleted sequences by subtracting the domains from the entire protein sequence. After the gathering of all the fragments, ARCA aligns the domains and the deleted sequences forming one multiple alignment. The generated architecture alignments showed that many of the deleted sequences between two domains from Pfam can be classified as domains by HMMER, showing new possible architectures.

Development a Predictor of Aggregation-protein using Supporting Vector Machine

Carlos Moreira, Prof. Dr. Luis Paulo Barbour Scott

Universidade Federal do ABC

Protein has important tasks for the body, such as the catalysis of chemical reactions, transport, and recognition and signal transmission. Although proteins could acquire a huge amount of conformations, they tend to a preferred conformation of lower energy known as native structure. However, there is a class of proteins that although it can be soluble in certain tissues it can be found as insoluble aggregates known as amyloid. According to several researches this kind of conformation is linked with diseases like: Alzheimer, Diabetes-II, and Parkinson. This proteins class coexists in two extremely different stable conformations: native and amyloid forms, the last one consisting mainly of β sheets. This work aims to develop a Predictor of Protein Aggregation concerning some features as: sequences of amino acids, physicochemical characteristics and trend to aggregation. The predictor will have a module considering energy frustration degree and secondary structures. For this purpose we investigated some tools described in literature like Aggrescan, Zyggregator, Page, Tango. We have been analyzing their algorithms and models adopted and the circumstances to use them. We are selecting and mapping some features of physicochemical characteristics, propensity of aggregation and statistical aspects of amino-acids. A local database is being created to store proteins amyloid, and other attributes. To set up the database we are using PostgreSQL tool. For the learning machine method we are using WEKA system/Support Vector Machine (SVM) at first. We are testing and making adjustments to reach higher performance for prediction. We are using Java language to program into the Linux environment. In the end the result will be compared with others predictors already mentioned in order to compare performance and accuracy.

Towards Transparent and Reproducible Bioinformatics Analyses: the EPIGEN-Brazil Scientific Workflow

Gilderlanio S Araújo¹, Wagner CS Magalhaes¹, Paula JS Viriato¹, Mauricio L Barreto^{2,3}, Bernardo L Horta⁴, M Fernanda Lima-Costa⁵, Alexandre C Pereira⁶, Eduardo Tarazona-Santos¹, Maíra R Rodrigues^{1*} and the Brazilian EPIGEN Consortium.

¹Department of General Biology, Institute of Biological Sciences, Federal University of Minas Gerais, Belo Horizonte, MG, Brazil; ²Instituto Gonçalo Muniz, Fundação Oswaldo Cruz, Salvador, BA, Brazil; ³Programa de Pós-Graduação em Saúde Coletiva, Universidade Federal da Bahia, Salvador, BA, Brazil; ⁴Programa de Pós-Graduação em Epidemiologia, Universidade Federal de Pelotas, RS, Brazil; ⁵Instituto de Pesquisa Rene Rachou, Fundação Oswaldo Cruz, Belo Horizonte, MG, Brazil; ⁶Instituto do Coração, Universidade de São Paulo, SP, Brazil

Two barriers in the dissemination and validation of knowledge today are lack of transparency and reproducibility of all the scientific process. It is argued that one way of tackling these issues is to have the complete scientific process as a research product in itself, which should be fully accessible and interactive. Following this direction, here we show how the EPIGEN-Brazil initiative website is being conceived as a Scientific Workflow and, at the same time, as a repository of bioinformatics resources for investigators in the areas of population genetics and genetic epidemiology. Both concepts are implemented as web tools and are fully and freely accessible online. Our Scientific Workflow is divided in self-contained components that cover different stages of the scientific process. In addition to Scientific Publications that comprise research hypotheses and results, a broad visualization of the research tasks is given by Flowcharts, the practical execution of such tasks is given by Masterscripts, and the intermediate steps, such as workshop discussions and technical reports, by Documents. A content manager connects conceptual and practical tasks and gathers comments from visitors and users. The latter feature allows user feedback to improve existing scientific analyses. We present examples in population genetics and genetic epidemiology that show how our Scientific Workflow can be intuitively navigated by researchers looking for guidance on similar projects or replication of existing results. Finally, we believe that our simple and intuitive approach to disclosing our scientific processes can encourage similar practices that increase the value of research.

LabControl: a LIMS software to manage microbial data

Mariana Teixeira Dornelles Parise¹, Joarley Ferreira dos Santos², Aristóteles Góes-Neto¹, Daniela Arruda Costa¹, Anne Cybele Pinto¹, Gabriel da Rocha Fernandes³,
Vasco Azevedo¹

Federal University of Minas Gerais¹, Pitágoras College², Research Center Rene Rachou³

Due to NGS sequencing techniques, the amount of genomic data has been increasing exponentially and this rapid growth has created management issues. Thus, Laboratory Information Management Systems (LIMS) have been used to manage this data and other types of laboratory data, such as samples and genomic data. Biological collections are important sources of raw material for genomic studies as well as essential ways to preserve the samples for future studies. In addition to these softwares, data patterns concerning genomic and microbiological fields have been developed and implemented in some LIMS or other kinds of laboratory softwares. Those patterns aim to standardize the information held in the softwares, facilitating information comparison and exchange between different sources or laboratories. However, neither the management utilizing softwares nor the usage of data patterns are common in small academic laboratories, which normally cannot afford for a commercial LIMS or develop their own LIMS. This lack of management support can cause inefficiency in the laboratory environment due to unorganized and unstandardized data in spreadsheets, data loss and difficult knowledge extraction and comparison. Considering this situation, LabControl has been developed as a free LIMS software based on World Federation for Culture Collections and Genomic Standards Consortium recommendations that aims to assist microbial collection management and organization by normalizing and integrating data from different organisms as well as to manage datasets and meta-data generated by *in vitro*, *in vivo* and *in silico* techniques applied to these strains. In addition to these features, the software requirements were based on interviews and investigation of researchers' routine, which generates a more reliable data model providing better assistance from the software to the researchers' real needs. The technologies used in the development are Java programming language in NetBeans IDE integrated with Spring and Hibernate frameworks, using the PostgreSQL database. LabControl software allows users to a faster strain information management through an easy-to-use system, which reduces the impact of the migration from spreadsheets to a web system. This migration process may promote a significant gain in researchers' daily productivity through organized, updated and persistent data of all studied strains and techniques applied to them.

Noninvasive prenatal paternity determination by SNPs and microhaplotypes

Jaqueleine Y. T. Wang^{1,2}; Renato D. Puga³; Martin R. Whittle⁴;

André Fujita^{2,5}; Helder T. I. Nakaya^{1,2}.

1 - School of Pharmaceutical Sciences - Department of Clinical Chemistry - University of São Paulo; 2 - Bioinformatics Graduate Program - University of São Paulo; 3 - Hospital Israelita Albert Einstein; 4 - Genomic Engenharia Molecular; 5 - Institute of Mathematics and Statistics - Department of Computer Science - University of São Paulo.

Invasive procedures like amniocentesis and chorionic villus sampling can be risky for pregnancy and may result in miscarriages. However, when information from the fetus is needed to investigate congenital abnormalities (such as sex-linked disorders and aneuploidy) or perform paternity tests, this risk is usually acceptable. Fortunately, the discovery of fetal DNA (fetal cell-free DNA, fcfDNA) in maternal plasma and the development of techniques to analyse this fcfDNA have allowed researchers to reduce this risk to fetus and mother. Microhaplotypes are chromosomal segments smaller than 200 bp, with at least three distinct haplotypes (alleles). Their average heterozygosity has to be greater than that of any of the Single Nucleotide Polymorphisms (SNPs) contained within them. Since the fcfDNA has a size of 166bp, it is sufficient to contain microhaplotypes which can be sequenced using Next Generation Sequencing (NGS) technology. The aim of this project is to determine the probability of paternity using SNPs within microhaplotypes. Raw sequencing data from three DNA samples are analysed: the alleged father, the mother and the maternal plasma (mixture of mother and fetus cell-free DNA). We performed sequencing quality control with FastQC and verified sequencing coverage with BEDtools. An optimum sequencing coverage is vital to infer the fetal fraction in the maternal plasma sample. Next, variant calling was performed with SAMtools, BCFtools and VCFtools. Finally, SNP annotation was done with ANNOVAR. All analysis steps were performed sequentially by using a pipeline written in Perl programming language. Microhaplotypes were chosen based on previous literature and human frequencies were calculated using data from 1000 Genomes. Combining genotype information, populational frequencies and fetal fractions, we will develop a method to calculate the probability of paternity in non-exclusion cases. Twenty microhaplotypes believed to be very informative for paternal investigation have been chosen based on their heterozygosity. The microhaplotypes frequencies have been calculated based on all the ethnic groups from 1000 genomes data. Currently, 30 sets of the three types of samples have been processed and annotated.

Novel bioinformatic approaches for viral discovery from NGS data

Liliane S. Oliveira¹, João M. P. Alves¹, Dolores U. Mehnert², Alan M. Durham³, Paolo M. A. Zanotto² and Arthur Gruber^{1*}

¹Dept. of Parasitology and ²Dept. of Microbiology, Institute of Biomedical Sciences, USP, São Paulo, Brazil;

³Dept. of Computer Sciences, Institute of Mathematics and Statistics, USP, São Paulo, Brazil.

*Correspondence: argruber@usp.br

Viruses are the most abundant biological entities and play an important role in defining the composition of microbial communities. Some of the most devastating pandemic diseases have arisen through the transmission of viruses originally infecting wild and domestic animals. Thus, a systematic surveillance for emerging viruses with new computational tools is of utmost importance. In this work, we report the development and implementation of some bioinformatic approaches using profile HMMs (pHMMs) for viral discovery. Profile HMMs are used in a variety of bioinformatic applications and the most relevant publicly available databases for viruses are vFam and viralOGs (a subset of eggNOG). We developed virDB-Pack, a suite of programs to quantify, manipulate and select pHMMs from vFams and viralOGs databases. The package allows using the selected pHMMs to screen sequencing datasets for known and potentially emergent viruses. A preliminary survey using 506 selected pHMMs against a metagenomic dataset from raw sewage revealed a variety of novel viruses, including sequences from smacovirus, densovirus and circovirus. Another dataset, composed of genomic reads of *Aedes aegypti*, allowed us to identify a large number of viral sequences integrated into the mosquito genome (see abstract by E. Aguiar & A. Gruber – X-Meeting 2016) and a 9-kb genome segment from a recently described Phasi Charoen Like-virus (PCLV). We are also developing an algorithm for the construction of pHMMs based on an iterative enrichment of viral sequence representation. The method involves an initial screen of the sequencing dataset with a pHMM constructed from a few sequences, and recruitment of positive reads and sequence reconstruction using GenSeed-HMM (Alves *et al.* - Front Microbiol. 7:269, 2016), a tool recently developed by our group. Resulting contigs are then translated and aligned, with the most conserved blocks being automatically selected and used to construct distance trees. Sequence and taxonomic redundancy are eliminated by clustering using patristic distance and user-defined parameters. The selected representative sequences are used to build a new pHMM that covers a higher viral sequence diversity. Preliminary results demonstrate that increasing the number of non-redundant sequences used to generate pHMMs leads to a better ability to detect viral diversity until a saturation point is attained. The approach proposed here, when applied to a wide number of viral families, will allow the detection of viruses phylogenetically distant from those already known, with potential application in viral discovery studies and epidemiological surveillance. Support: Fellowships from CNPq and CAPES.

Software Assessment for Prediction of Gene Clusters: An Analysis *in silico* with Cyanobacteria of Chroococcales Order

Danielle C. C. Couto, Vanessa C. Rezende, Alex R. J. Lima, Felipe C. Couto,

Leonardo T. Dall'Agnol, Evonnildo C. Gonçalves

Federal University of Pará, Federal University of South and Southeast of Pará, State

University of Pará, Federal University of Maranhão

The main objective of this study was the comparison of four different cluster gene prediction tools available (antiSMASH; NP.searcher; NaPDoS; DoBISCUIT) and the influence of the input of the biological information (fasta; annotated; etc). Our work compared the tools using three cyanobacterial genomes from Chroococcales order: *Cyanobium* CACIAM 14, *Synechocystis* PCC6803 and *Synechocystis* CACIAM 05. The results showed that the integration of the generated data between the different prediction tools promotes deeper and better prospection of clusters. It is important to highlight that depending on the data input format directly influences the number of groups detected, helping to unravel the biotechnological potential of the organisms. The online antiSMASH 3.0.2 in *Cyanobium* sp. CACIAM 14 varied the number of predicted clusters according to the imput: there were 24 clusters for RAST entry against 35 for fasta. Moreover, it is noted that the online version can predict more clusters, due to the incorporation of saccharides and fatty acids clusters. The genome of *Cyanobium* sp. CACIAM 14 was also run on antiSMASH local version 2.0 using gbk and fasta files, annotated with the NCBI PGAP and RAST. Both have generated a total of 15 clusters for gbk and 14 for fasta. Online antiSMASH analysis of *Synechocystis* sp. CACIAM 05 generated as a result 23 clusters for fasta and 16 for gbk. antiSMASH online screening of *Synechocystis* sp. PCC 6803 revealed 33 cluster for fasta and 11 for gbk. However, the result may vary greatly depending on the input type and the types of software used for the genome annotation. The results of NP.searcher tool were more limited, as these cyanobacteria have few (rarely) NRPS/PKS modules. The NaPDoS also had limited results, but has interesting features such as the fact of presenting a tree with the expected product structure, the BLAST results and candidate domains. The DoBISCUIT tool database is extremely extensive and so has a great biotechnological potential related to the prospection of bioactive products in cyanobacteria. Due to this, it presented the highest number of predicted clusters. The mains result of this work was to prove that there are important differences among the results of different gene cluster predictive tools according to the input information and that different tools and parameters should be combined to avoid enlarge the results.

Brimer: A Web System for Managing Primers

Danielle C. C. Couto, Aryane P. Vilhena, Felipe C. Couto, Leonardo T. Dall'Agnol,
Hivana P. M. B. Dall'Agnol, Evonnildo C. Gonçalves

Federal University of Pará, Federal University of South and Southeast of Pará, State

University of Pará, Federal University of Maranhão

BRIMER - Primer Library Online is a web software implemented for better management of the oligonucleotide library of the Biomolecular Technology Laboratory (LTB) researchers from Federal University of Pará. Primers are small single stranded DNA chain, which binds to a specific region of the target, wherein the DNA polymerase will begin incorporation of nucleotides forming the double-stranded DNA. This work aimed to develop a web system for primers catalog management with user access for LTB personnel. There already exists some oligonucleotide management software available in the literature; PIPEMicroDB is the more similar with the same programming language (PHP), online relational database and the same database management system (MySQL). However, the PIPEMicroDB is a microsatellite database and primer generation tool for pigeonpea genome; Brimmer, on the other way, can be used to build an organized catalog of primer sequences of any organism. This database would generate a library of primers with all relevant information for better managements as: (i) related scientific communications/description; (ii) synthesis costs; (iii) target genes and organisms; (iv) any other useful information. The Brimer system is hosted initially in <http://ltb.ufpa.br/brimer>, which features presents in its initial page, user login and password. After accessing the system, the user authentication is validated and Brimer verify its access level and profile. A Primer query screen can be seen by all profiles and offers three options of search filter: "Oligo", "Organism" and "Gene". If necessary, the system also exhibits other registered primers in the database that present the same melting temperature for optimizing thermocyclers usage. Additionally, the system offers the functionality of linking PDF files of bibliographic items related to each primer. The Brimer system was implemented and tested on Windows, Linux and Mac OS; Chrome and Mozilla Firefox web browsers. This work applied the Kanban's agile methodology development techniques to implement a prototype for testing and validation of the system. We used a qualitative methodology involving documentary and bibliographic research, field assessment at the LTB group to evaluate the system requirements through interviews and checklists. Brimer was tested by LTB personnel and the user evaluation was measured by questionnaires applied individually. The survey results indicate a good acceptance and satisfaction from users who are using the system, with the better evaluated feature being the system ease of use. Researchers also pointed that Brimer is very important to optimize the organization and utilization of the oligonucleotide currently used at the laboratory.

Network Algorithm To Relatedness Analysis (NAToRA)

Thiago Peixoto Leal, Mateus Gouveia, Gilderlanio Santana de Araújo, Maíra R Rodrigues, Marilia Scliar, Eduardo Martin Tarazona Santos

Laboratório de Divergência Genética Humana (LDGH) , Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais

The relatedness can cause population stratification resulting in spurious associations in genome-wide association studies (GWAS) and biases on population structure studies. To correct this problem there are several methods that estimate the relatedness of the samples (Purcell et al 2007 , Thornton et al 2012), giving IBD-sharing probabilities (Identity by descent, ie, the probability that a matching segment of DNA shared by two or more individuals has been inherited from a recent common ancestor). The statistics are composed of pair-wise IBD0, IBD1 and IBD2, which are probability of not share alleles by descent, share one allele by descent and share two alleles by descent respectively. Through IBDs we can calculate the kinship coefficient between two individuals i and j (Φ_{ij}) by the equation $\Phi_{ij} = \frac{1}{4} \text{IBD1}_{ij} + \frac{1}{2} \text{IBD2}_{ij}$. The theoretical values of Φ_{ij} that correspond to the following degree of relatedness between i and j are: 0.5 for self or twins, 0.25 for first degree, 0.125 for second degree, 0.0625 for third degree, 0.03125 for fourth degree and 0 for unrelated. Although there are several methods of estimating kinship , there is no method in the literature that tries to create a population unrelated sample trying to minimize the exclusion of individuals. Using the Graph and Complex Network Theories, NAToRA detects the families in the network and eliminates individuals to create a sample without kinship. The first step is to create a Network (a Network (Network N) where the nodes are the individuals and the edges are the kinship coefficient between the nodes. After this step, eliminate all edges between nodes with a value lower than a cut-off value α , ie, what degree of relatedness to be considered (Network N_c). Each cluster in N_c (Connected Component) is a Network Family. As the problem of obtaining a smaller number of individuals to be removed to create a free edge network is a NP-Complete problem, we have implemented a heuristic. The algorithm calculates the node degree centrality and remove the highest degree (ie, the individual with more relatives) until only exist pairs of individuals and edgeless nodes. With the pairs we look for the Network N , calculate the centrality for both nodes and exclude the highest. After the process, the algorithm gives a list of families and the individuals to remove to create a subset without relatives.

HaploCYP: a software for *CYP2D6* genotyping and phenotype prediction

Pereira MA¹, Cardenas RGCL², Mateo ECC¹, Ferreira ACS¹, Freire MCM¹

¹*Setor de Pesquisa e Desenvolvimento, Grupo Hermes Pardini, Vespasiano, Brasil;*

²*Torchmed - Software Development, Belo Horizonte, Brasil; ³Laboratório Progenética, Grupo Hermes Pardini, Rio de Janeiro, Brasil*

Cytochrome P450 2D6 (*CYP2D6*) gene (MIM#124030) is one of the most polymorphic pharmacogenes with more than 109 allelic variants reported up to date by the Human Cytochrome P450 (*CYP*) Allele Nomenclature Database. The polymorphisms include SNPs, indels, copy number variations (CNVs), conversions and gene rearrangements. Due to this genetic polymorphism, *CYP2D6* exhibit notable inter-individual variability in enzyme activity and individuals can be divided into four phenotypic groups: poor (PM), intermediate (IM), extensive (EM) and ultrarapid (UM) metabolizers. Since *CYP2D6* genotype assignment and phenotype prediction are complex and of utmost importance into clinical practice, this work aimed to develop and validate a user-friendly software for *CYP2D6* genotyping and phenotype prediction using Sanger sequencing and CNV data. HaploCYP combines a set of python modules, BLAST tool and a MySQL database system in a web interface running on an Apache web server. The workflow consists of: BLAST alignment, polymorphism detections, genotype annotation and phenotype prediction. CNV information and fasta files from Sanger sequencing of *CYP2D6* gene are given as input. Variants are detected through BLASTn alignment with *CYP2D6*1* reference sequence (Accession Number: AY545216.1). Mutation nomenclature and haplotypes are defined according to *CYP* Allele Nomenclature Committee, PharmGKB and LOVD databases. The haplotypes that best represents the set of polymorphisms and CNV information are defined and reported following the star-allele nomenclature system. In some cases, more than two haplotypes can be reported. At this point, the user needs to review the haplotypes proposed and choose the correct genotype. Then, the phenotype is predicted as PM, IM, EM or UM, based on *CYP2D6* diplotypes and the activity score system recommend by the Clinical Pharmacogenetics Implementation Consortium (CPIC) guidelines for codeine therapy. Validation included simulation (sequences from NCBI) and ten real data with known *CYP2D6* genotypes. HaploCYP was able to genotype correctly all the simulation datasets and nine real data. For one sample, the program was not capable to do genotyping annotation since two genotypes were equally likely. However, the expertise of the user was enough to solve this genotyping problem. A higher dataset will be used to test accuracy, specificity and sensibility. HaploCYP simplify and facilitate the genotyping and phenotype prediction process of *CYP2D6* into clinical diagnosis, where speed and precision is of high importance. In addition, HaploCYP can be used for others *CYP* locus and any pharmacogenes that it is necessary to identify pharmacogenomic variants to individualize drug prescription and analyze drug efficacy and safety.

Financial Support: Hermes Pardini Group, FINEP

Text mining for HPC

Bruna Piereck Moura¹, Adriano Barbosa da Silva², Ana Maria Benko-Iseppon¹ e
Ana Christina Brasileiro Vidal¹

¹*Universidade Federal de Pernambuco - PPGG*, ²*Luxembourg University, Luxembourg - LCSB*

Each day, the volume of published data in biomedical and biological research is exponentially increasing, becoming a challenge to keep up dated the knowledge about a given topic of study. PubMed/Medline had approximately 22.7 mi citations until 2014/2015 and it has accomplished around to 26 mi citations until September, 2016. PESCADOR (Platform for Exploration of Significant Concepts Associated to co-Ocurrence Relationship) is an online flexible text mining tool that aggregate other programs to identify molecules pair of interactions. The objective of this work was to improve PESCADOR, turning possible HPC usage to analysing more data in less time. To achieve that, the tools and processes used by PESCADOR were individually adapted to HPC environment. At first, the 779 xml files from Medline database composed by approximately 30.000 citations each were automatically parsed to recover the PMID, Title and Abstract using python script. Followed by NLProt tagging tool, to highlight the protein and DNA names on xml parsing output (txt file). Analysing 10 files at time on one node (12 CPUs, 2 cores each) using the *parallel* tool, with 12 nodes-job, letting 2 free CPUs on each node. At last, based on MySQL and written in PHP language the LAITOR program, responsible for de interaction and co-occurrences identification, was modified, using python script, to SQLite format becoming able to run on HPC environment, all the libraries and index as much the query PHP lines at the original program were up date. LAITOR was run on one node, 8 files at time using *parallel* tool, letting 4 free CPUs. The xml parsing was finished after around 1h and was followed by the NLProt tagging, that was running for 5 days. At Last the LAITOR was running for around 3-4 days. The results of more than 23 mi abstracts were complete to be statistically analysed and curated. The same would not be possible in the online version of PESCADOR. All the adaptation needed took six (6) months. This results statistics and curation will make possible to enrich MESH terms, evaluate the well and poorly described protein interactions and estimate the needed time to curate all the interactions described until now.

Financial support: CAPES, CNPq, FACEPE

Database Model for Fish Collection

Nicolás Valentín Molina Terra, Maria Fernanda Hussni, Luiz Henrique Garcia

Pereira, Marcelo Cezar Pinto

UNILA – Universidade Federal da Integração Latino-Americana

Biological data collection is an important activity used by Botany, Ecology and Geographical Analysis research fields. Biological data acquired in a field collection trip at some site can provide much significant information in a scientific project and the data availability is important to the collectors as well as to the research community. So it is a great benefit to them if these data can be accessible by a web database that is modeled to fit the required acquisition process used by researchers. Besides the data input process it is also important to have a good query system and to allow the addition of analysis plugins. This work in progress project aims to devise a complete website for the Fish Collection of UNILA, also integrating a Botany database and a Geographic Information System. At this point of the project we have made the database modeling and a prototype of the data input web interface. These activities were made by a multidisciplinary group of mathematicians, biologists and computer scientists. An expert researcher from Ichthyology was interviewed and provided the required information to the database system as well as the web interface for fish batches and tissues collections. Besides that, some batches and tissues loan system was sketched to allow a borrowing log of biological material. So far we have used HTML5, CSS and JavaScript languages for the web client side of our Fish Collection and the KORA framework to a fast implementation of the database system. Temporarily we are hosting the Fish Collection at <http://54.227.226.95/kora-2.6.6.1/> (login required). This project is registered as PID202-2015, PID495-2016 and PID575-2016 and is partially funded by PIBIC-UNILA and PIBITI-UNILA.

A comprehensive database of mirtrons knowledge

Bruno Henrique Ribeiro da Fonseca¹, Douglas Silva Domingues^{1,2} and Alexandre Rossi Paschoal^{1*}

¹ Programa de Pós-Graduação em Bioinformática - PPGBIOINFO, Universidade Tecnológica Federal do Paraná, Cornélio Procópio, ² Departamento de Botânica, Instituto de Biociências, Universidade Estadual Paulista, Rio Claro

* Corresponding author: paschoal@utfpr.edu.br

MicroRNAs (miRNAs) are one class of small non-coding RNAs (ncRNAs) that occurs in several eukaryotic genomes. They are responsible for post-transcriptional control of mRNA levels in cells. Recently, mirtrons were discovered as an alternative miRNA class, whose biogenesis is based on the splicing process as the first cleavage step in miRNA maturation. Up to now, most public data in mirtrons are restricted to few model organisms: *Arabidopsis thaliana*, *Oryza sativa*, *Homo sapiens*, *Mus musculus*, *Drosophila melanogaster*, and *Caenorhabditis elegans*. These data are available in many different sources and structures. To our knowledge, there is no central repository or integrated web resource for mirtron research. In order to fill this gap, we describe mirtronDB, a public repository for mirtrons. This database has as major contributions: (i) integration of all public data available in literature under a unique and structured database and system, and (ii) become the first friendly central repository about mirtrons. Among the functionalities, we highlight: (a) general search (e.g. names, identification); (b) similarity search; (c) graphic visualization; (d) conservational analysis among organisms; and (e) comparative analysis between mirtrons and canonical miRNA. This web application will be built using PHP language, with Chado scheme, in a MySQL database. For that, we: (i) analyzed and extracted all public data in the literature about mirtrons; (ii) organized and structured all data for these organisms; and (iii) we will built a web application. We think that this repository will allow and facilitate the scientific communities apply methods and computational techniques in bioinformatics in this data.

Support: CNPq (#454505/2014-0).

Linking microbial community composition and potential functional roles using shotgun metagenomic libraries

Laura Rabelo Leite¹, Julliane Dutra Medeiros^{1,2}, Francislon Silva de Oliveira^{1,2}, Victor Satler Pylro¹, Sara Cuadros Orellana¹, Guilherme Corrêa de Oliveira³, Gabriel da Rocha Fernandes¹

Centro de Pesquisas Rene Rachou – Fiocruz Minas¹, Universidade Federal de Minas Gerais² and Vale Institute of Technology – Biodiversity and Biotechnology³

Metagenomics involves the study of genetic material recovered directly from environmental sample, allowing microbiologists to analyze not cultivable organisms. Shotgun metagenomic reads can be taxonomically or functionally classified. Classification methods can be divided in three strategies: (a) sequence similarity methods, which use the results of a sequence similarity search against a database of a reference set of sequences, (b) sequence composition methods, which are based on characteristics of their nucleotide composition, (c) marker-based methods which identify species based on the occurrence of specific marker sequences. However, the accurate identification of microorganisms at the species level remains extremely challenging. Here, we proposed a pipeline to analyze, in house, shotgun metagenomic reads (classifying them taxonomically and functionally), resulting in a relational database to link these information. Raw reads were subject to quality filtering using Trimmomatic. A taxonomic profile was obtained by comparison of all reads against the NT database, and the Nucleotide BLAST output file was used to calculate a taxonomic classification based on the lowest common ancestor method. Contigs larger than 500bp were obtained with SPADES assembler. Coding sequences were predicted with MetaGeneMark. Proteins were extracted using an in-house Perl script and classified into UniRef Enriched KEGG Orthology using DIAMOND. The KO counts were normalized according to the length of the target gene. Reads were then mapped back onto the contigs, using Bowtie2, to determine the contribution of each taxon to the environment gene pool. Our pipeline allows the inference of direct relations between taxonomic and functional data in high-throughput libraries. Moreover, is possible compare the relative abundance of organisms and metabolic pathways between environments and look for taxa that carry out functions of interest. Our pipeline was evaluated against a mock community metagenome, downloaded from Human Microbiome Project (SRR172903), and compared to two approaches: MetaPhlAn2 (based on marker genes) and GSM (based on k-mer composition). The sensitivity evaluation suggested that our pipeline was sufficient to identify microbial strains with $\geq 0.2\times$ coverage, and 63% of selected genus should be detected based on the positive predictive value. Still, GSM showed better specificity results (51%), due to low false positive rate, highlighting the indicator that needs to be improved in our pipeline. The proposed pipeline was also used to build the first catalog of the biotechnological potential of microorganisms in brazilian copper mines, identifying about 3,800 potential commercial applications related to CUEs (Commercially Useful Enzymes).

DevOps cloud-computing environment to perform virtual screening

Léo Rodrigues Biscassi¹, Rodrigo Antônio Faccioli², Paulo Eduardo Ambrósio¹

Universidade Estadual de Santa Cruz¹, Centro Universitário Barão de Mauá²

Failures in all phases of clinical trials have been increased the past decades. It is occurring despite improvements in all stages of the drug development pipeline. One of the key areas of improvement has been the virtual screening for drugs likely to fail clinical trials. The drug-likeness measures have been widely accepted as a useful guide to identification of promising molecules to be tested as drug candidates in the early stages of drug discovery. As consequence of virtual screening is the focus on efforts and resources on experiments with promising molecules. The virtual screening technique demands a dynamic and heterogeneous computational environments which can adapt itself according to each task. It is a challenge to drug discovery projects that have used for along the time. In order to attend this challenge is employed DevOps. It means the practice lays emphasis on the collaboration and communication between software developers and professionals while automating the process of software delivery and infrastructure changes. Resulting, DevOps aims at establishing a culture and environment where building, testing, and releasing software can happen rapidly, frequently, and more reliably. In this work is presented a DevOps strategy for virtual screening through docker and Galaxy project. Galaxy project is an open-source web-based platform what made easy the reproducible research and provides a good engine to make friendly interfaces to command line tools. Docker is an open-source platform which consists in docker container engine and docker hub. The docker container engine is responsible to create and manage isolated containers on top linux kernel with his technologies called namespaces and cgroups. The docker hub makes possible deploy and share our own images on docker environment with other people. We've implemented a docker image which has a galaxy platform instance with tools to perform virtual screening with autodock vina allowing system administrators deploy and increase resources on demand in cloud environments. Furthermore it allows the users have a user-friendly interface to autodock vina and track all steps made in your analysis. In some tests accomplished for us the time to get up the container working was on average 5 minutes, which is a good result compared to the 2 hours on average taken with the same tests performed manually.

BDGF: a database and web-based information retrieval system for genotype and phenotype

Fábio Danilo Vieira, Danilo Gomes de Moura, Diego Félix da Silva, Roberto

Hiroshi Higa, Adhemar Zerlotini

Embrapa Agricultural Informatics, Campinas, SP, Brazil

In recent years, the use of large scale genotyping of tens or hundreds of thousands of Single Nucleotide Polymorphisms (SNPs) to estimate the genomic profile allowed the development of both genotype-phenotype association studies in genomic scale (genome-wide association studies - GWAS) and the introduction of genomic selection technology in breeding programs. However, this situation implies the need of storing large volumes of genotyping, phenotyping and pedigree data from large numbers of animals, a trend that will likely increase over the coming years, given the lower costs for generating the experimental data. In order to effectively integrate such amount of distinctive datasets, it's advisable to use a robust storage structure, such as a DBMS. Therefore, a major issue to consider is the trade off between normalization and performance during the database modeling stage, as this will have a direct impact on the usability and user experience. In order to get efficient storage and fast queries in this high volume of data, in this work we present the BDGF system (Genotypes and Phenotypes Database). It is based on a data model first proposed by (HIGA, 2015). Nowadays, noSQL has vastly improved our capacity to handle big data, and became integral part of traditional DBMS, such as PosgreSQL. BDGF was completely remodeled so that it has advantages in the use of such technologies. The JSON technology is widely employed in order to allow flexibility to store any phenotype and guarantee immediate query results regardless the number of records. BDGF is designed to support the animal breeding projects of Embrapa, but can be easily adjusted to store data from diverse sources, such as clinical or plant data. Furthermore, the system implements access and security policies to phenotypes, genotypes and pedigree of the animals. The system was developed using webstandards, i18n and free software tools, such as Java, Primefaces, Hibernate and Jboss. BDGF is currently being documented and tested and it's expected to be fully operational within a year.

PFStats: A tool for protein analysis by decomposition of residue coevolution networks and amino acid reduced alphabets applications

Neli Fonseca, Lucas Bleicher, Marcelo Querino

Universidade Federal de Minas Gerais

Structural and functional insights about protein families can be obtained by amino acid conservation and correlation analysis. Furthermore, experimental research has suggested that protein folding can be achieved with fewer characters than the 20 naturally occurring amino acids. Our group has recently proposed a method to obtain functional sub-class determinants in protein families, called Decomposition of Residue Coevolution Networks (DRCN). DRCN is a sequence based method for analysis of protein families represented by multiple sequence alignments. We present a software for protein family analysis using DRCN, conservation analysis, alphabet reductions, and automatic annotation search. The algorithms were grouped in order to have a robust and intuitive application to the analysis of homologous proteins. The DRCN analysis consists of a unique required input file, a multiple sequence alignment (MSA), besides that a PDB file can be also used to visualize the results in the structure. The MSA quality is a crucial factor to achieve better results with the methodology, therefore, a filtering step is available to maximize its representativeness by removing fragments, poorly aligned sequences and redundancy. We have studied four protein family domains: lysozyme C/Alpha-lactalbumin, phospholipases A2, nitrogen regulatory protein PII, and the DNA binding domain of the nuclear receptors IV; three MSAs approaches extracted from Pfam and 19 amino acids reduced alphabets from literature. We have found insights about catalytic and binding sites in all of them. There's also information related to secondary structure, the hydrophobic putative channel, and dimerization sites. By looking for the anti-correlated edges, we could find a residue or a group of residues that separates two or more sub-classes. That's the case of the C122 in the phospholipase A2, this node forms an anti-correlated hub that connects every community. Its presence occurs in 217 sequences, all from *Oikopleura dioica*, and all without the phospholipase catalytic activity. The uses of reduced alphabet in DRCN analysis usually increase the number of residues in each community and in the most cases maintaining a consistent hypothesis for their biological role. But in these cases, nuclear receptors IV study, the uses of a reduced alphabet can hide clusters that share common positions with another community.

Funded by FAPEMIG and CAPES

Classifiers for patients with breast cancer according to the neoadjuvant chemotherapy sensitivity

Pedro Kássio R. M. L. Carvalho, Thiago de Souza Rodrigues

CEFET-MG, CEFET-MG

The breast cancer is the most common tumor on women, about 508000 died in 2011 due to this disease. The treatment is usually done with neoadjuvant chemotherapy, followed by the operation to remove the tumor and after, the adjuvant chemotherapy. The neoadjuvant chemotherapy may show Complete Pathological Response (PCR), when the disease is completely eliminated, or, on the other hand, Residual Disease (RD). This project uses information about the molecular subtypes of breast cancer in order to classify the patients according to the chemotherapy sensitivity. Among the subtypes, the basal-like was not used, because of its difficult in classification problems. A dataset composed by gene expression of the patients, extracted from Gene Expression Omnibus repository, was used to create classifiers based on machine learning techniques, computational intelligence and evolutionary computation. Feature selection methods were applied in order to select the best characteristics to create the classifiers. From univariate feature selection method Volcano Plot, we selected 31 genes. From the multivariate feature selection method stepwise regression we selected 110 genes. And from the regression based on the Generalized Linear Model we selected 186 genes. Classifiers were created using different algorithms and the filtered data base. Six using neural networks with different types of training algorithms, one with particle swarm optimization with clustering and one with extreme learning machine algorithms. The neural network classifiers presented an average result of 52% accuracy. With the particle swarm optimization the best result was 62% of accuracy, using the 186 genic expressions. The best classifier was obtained using the Extreme Learning Machine algorithm, which has a very small runtime and 80% of accuracy on average, indicating a good result, which must also be adjusted to improve the hit rate. The genic expressions that showed this result were the 186. We can see that the extreme learning machine appears to be the most appropriate algorithm found for this problem and has the best runtime and results. It can still be improved to get a better result using fewer genic expressions, but has already shown a good initial result. This work is supported by CAPES, CNPq and FAPEMIG.

Impact of genomic RNA structure and non-coding RNAs in Zika virus neuropathogenesis

Raúl Arias-Carrasco, Yessenia Vásquez-Morán, Fernanda Castro, Artur L. de Queiroz, Helder I. Nakaya, Renato S. Aguiar, Vinicius Maracaja-Coutinho

Centro de Genómica y Bioinformática (Universidad Mayor – Santiago/Chile), Departamento de Genética (Universidade Federal do Rio de Janeiro – Rio de Janeiro/Brazil), Gonçalo Moniz Research Center (FIOCRUZ – Salvador/Brazil), Faculdade de Ciências Farmacêuticas (Universidade de São Paulo – São Paulo/Brazil), Instituto Vandique (João Pessoa/Brasil), Beagle Bioinformatics (Santiago/Chile)

Zika Virus (ZIKV) infection has been neglected during the first 60 years after its discovery. In the last years the unexpected rapid spread and potential ability to cause congenital syndrome have made ZIKV a Public Health Emergency of International Concern. ZIKV genome has two flanking UTR regions and a single long open reading frame encoding a polyprotein. In other Flavivirus, the 3'UTR sequence form particular RNA structures, which are important for virus biology and lifecycle. It is unclear, however, if variations in these viral secondary structures may also contribute to the neuropathogenesis induced by Zika infection. In this work, we developed and applied an automated pipeline called StructRNA under on 55 publicly available 3'UTR sequences from ZIKV and other 32 related aviviruses. Our tool provides an easy-to-use method to functionally annotate RNA structures nucleotide sequences (DNA or RNA). The same set of sequences was also submitted to analysis on PPfold software, in order to find the differences in terms of pairing composition and general topology of ZIKV 3'UTRs. Our analysis found a group of 47 ZIKV exclusive RNA families; and other 25 families available exclusively on other aviviruses. The results revealed differences in terms of secondary structure composition along this region in different ZIKV isolates, and was able to group the sequences between Asian/American and African lineages. The PPfold analysis also illustrated clear differences between these two lineages, revealing that African lineage seems to be much more structured than the Asian lineage. These structural ZIKV RNAs, as well as the topology of the 3'UTR region, may have critical roles in virus biology, and can be directly related to human neurological disorders.

Financial support: CAPES/CNPq/FAPERJ/FAPESP/Santander Universidades.

DNAShot: an application to Blast DNA Sequence from photos using smart-phone

Ricardo Voyceik*, Roberto Tadeu Raittz and Fabio de Oliveira Pedrosa

Universidade Federal de Minas Gerais (UFMG), Universidade Federal do Paraná (UFPR)

The growing use of smart-phones in the most common tasks enhances ever more due to the increased availability of applications developed for these devices. In Bioinformatics this trend is not different, due the emergence of applications to process and display biological information for smart-phones and tablets. However, the ability to interact with non-responsive sites with these devices is limited to the use of native browsers. The DNAShot app is a new smart-phone application that recognizes DNA sequences from images. DNAShot is an iOS and Android OS based application. An OCR module converts pictures taken by built-in camera to a text format and then submits it to the NCBI site for Blastn task within nr/nt database and the top 10 hits are shown in a result list. This feature is useful in tasks like to determine the species of the organisms whose nucleotide data are not available to be uploaded or pasted in a browser. DNAShot can also handle stored DNA sequences images from memory card and/or from the Cloud. The search using the DNA Shot are accurate, as the tests showed when recovering known sequences from the NCBI database. The response time was also acceptable, once it depends mainly on the processing capability of the mobile device to convert the image and the availability of the Blast NCBI web service. DNAShot app brings mobility and efficiency to researchers since it can be used anywhere, anytime, as long as there is an Internet connection, offering new ways to support Bioinformatics and Biology researchers in a wide range of new situations. The authors thank to partnership between UFPR and UFMG Bioinformatics Graduate Programs and the financial support of Brazilian funding agency (CAPES).

Evaluation of predictor programs of genomic islands

Antonio Camilo da Silva Filho, Diônata Willian Augusto, Izabella Castilhos Ribeiro dos Santos Weiss, Paulo Afonso Bracarense Costa, Jeroniza Nunes Marchaukoski

Universidade Federal do Paraná (UFPR), Setor de Educação Profissional e Tecnológica, Programa de Pós Graduação em Bioinformática, Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq)

Genomic islands (GI) are segments of DNA characterized by gene content and are associated with a function or adaptive capacity of medical and environmental interest, being directly related to bacterial evolution. The GIs are present in different taxonomic clades and are often obtained by horizontal gene transfer. The classification of a GI is given accordingly to the function of encoded genes, it can be classified as pathogenic, symbiotic, metabolic, saprophagous, ecological or resistance islands. Among these, stands out the pathogenicity islands (PAIs) that has the ability to transmit pathogenic properties in a single genetic event thus allowing the evolution and the emergence of new pathogens. In this work, we evaluated predictors programs of genomic islands using only complete genomes. We analyzed the similarity between the results generated by the predictive programs and the gold standard classification. For testing programs we selected: Alien Hunter, Zisland Explorer, IslandViewer3, GI Hunter, Predict Bias and Gipsy. It was considered tree points of criteria to choose the predictors: 1- the availability (the predictor is local or web); 2 – article published less than three years and 3 - that has more than 20 citations in the last three years. We chose Escherichia coli CFT073 as a test organism. For the gold standard we consider the published pathogenic islands of Escherichia coli CFT073 for that it was researched and proved *in vitro* and this class of pathogenic islands is the most studied. In the analysis we observed differences in the results of the compared predictors to the 13 pathogenic islands which was determined as gold standard. Determination of these islands was analyzed *in vitro* in previous research. Each predictor showed different results, among them are the start and end positions of the island, size, number of coding sequences and the number of predicted islands. Alien Hunter identified 13 PAIs of the gold standard in 86 GIs predicted, IslandViewer3 identified 11 PAIs of 78, Predict Bias identified 11 PAIs of 76, Gipsy identified 12 PAIs of 26, GI Hunter identified 7 PAIs of 18 and Zisland Explorer identified 3 PAIs out of 11GIs. In order to evaluate the differences between the results, we need to identify what tools and methodologies have high and low sensitivity, specificity and accuracy determining the positive and false negative results, as well as real positive and negative islands. This research will help to find the best strategies to prediction, being constructive and direct for researchers to obtain promising results in studies related to genomic islands.

miRNAPath: platform to identify miRNAs targets and pathways regulated by miRNAs

Natália Baptista Cruz^{1,2}, Jessica Rodrigues Plaça^{2,3}, Wilson Araújo da Silva Jr^{2,4}

¹ Biomedical Informatics Graduation Program, Ribeirão Preto Medical School, São Paulo, Brazil; ²National Institute of Science and Technology in Stem Cell and Cell Therapy and Center for Cell-Based Therapy, Ribeirão Preto, São Paulo, Brazil; ³Clinical Oncology, Stem Cell and Cell Therapy Program, Ribeirão Preto Medical School, Ribeirão Preto, São Paulo, Brazil; ⁴Department of Genetics at Ribeirão Preto Medical School, and Center for Integrative System Biology (CISBi-NAP/USP), University of São Paulo, Ribeirão Preto, São Paulo, Brazil

MicroRNAs (miRNAs) are small non-coding RNA molecules that regulate their target messenger RNA (mRNA) at post-transcriptional level, altering important biological processes as apoptosis and cell proliferation. miRNA expression alteration has proved to be effective in modulating the passage signal through certain pathways involved in pathological processes, restoring its normal behavior in diseases. Although this interaction is well studied, there are few tools that allow the visualization of miRNAs/mRNAs expression correlation in a non intuitive way or tools that allow the user to provide his own expression data file from which the correlation analysis will be calculated. Thus, the aim of this study was to develop miRNAPath II, a web tool that calculates miRNA/mRNA expression correlation and indicates the possible signaling pathways involved in this interaction. For this project, the used data was obtained from public and freely available database, such as miRBase, KEGG, RefSeq and TCGA. This tool can perform two different analyses depending on what information the user is interested in. The first analysis evaluates the correlation between the queried miRNAs and their respective associated mRNAs previously described in the literature, followed by a differential expression analysis, which will display the target genes in a plot-formatted result. The second analysis will provide the pathways that are related somehow with the input miRNAs. Focusing on the first analysis, the user can evaluate the association between the input miRNA and mRNAs by providing a personal expression data file or by choosing an already available expression data in the TCGA database, allowing the user to choose the disease of interest. miRNAPath II was evaluated using expression data obtained from TCGA patients with colon adenocarcinoma and based on bibliographic research. The miRNAs hsa-mir-20a and hsa-mir-21 were selected as input. It were reported 164 mRNAs negatively correlated with the expression of the input miRNAs. Then, LGALS3, the first significant mRNA shown as result, was chosen to run an analysis of differentially expressed gene, indicating that this mRNA has indicative that it is a possible target for future treatments related to

A graphical tool for data integration and analysis of complex diseases

João C. Pandolfi¹, Ana Rúbia R. Vicente¹, David C. Martins-Jr², Sérgio N. Simões¹

¹Federal Institute of Espírito Santo, ²Federal University of ABC

Complex diseases are polygenic and multifactorial. For this reason, prioritization of genes related to complex disease is a challenge. Besides, one of the main problems for researchers is the lack of reproducibility among studies involving different methodologies or experiments. Particularly, Protein-Protein Interaction (PPI) networks have been used to prioritize genes related to complex diseases according to their topological features. Recent methods for prioritizing genes of complex diseases generally perform integration of expression data and PPI (Protein Protein Interaction) networks. Most methods exploiting PPI networks are based on Network Medicine hypotheses, from which we highlight: network parsimony, locality and disease module. A method called NERI (*NEtwork-medicine based integrative approach for disease gene prioritization by Relative Importance*) recently published prioritizes genes by exploring PPI network based on Network Medicine hypotheses and relative importance algorithms. Despite having shown good results, currently the implementation of this method lacks a graphical user interface, and provides only a command line interface, thus hindering their use by researchers. In this work, we developed a graphical user interface for NERI, as well as ETL (*Extraction, Transform, Load*) of biological data in various formats. In this system, the user provide as input the expression data, PPI and GWAS (genes used as seeds). These inputs are facilitated by the graphical interface, which then calls the processing method NERI giving as result the prioritized genes and complex networks differential analysis. Currently, this interface is already functional, and improvements are currently being made in visualization of complex networks. Thus, the graphical tool developed in this work facilitates the application of this methodology by the end researcher. As future work, we intend to provide the system as a web service and free software.

Identifying Alternative Splicing Events in RNAseq data using De Bruijn Graphs and Bloom Filters

Ricardo Medeiros da Costa Junior, André Yoshiaki Kashiwabara

Universidade Tecnológica Federal do Paraná - Campus Cornélio Procópio, Departamento de Computação, Programa de Pós-Graduação em Bioinformática - PPGBIOINFO

Alternative splicing (AS) is a post-transcriptional mechanism in which multiple functional transcripts might be produced from a single gene. In particular, a gene encoding the protein may produce different proteins through pre-mRNA AS events. In this process, some exons may be included or excluded from the final messenger RNA (mRNA). In consequence, AS mRNA translated protein contains differences in their amino acid sequences and often in their biological functions. The AS process allows the human genome directly synthesize many proteins that could be expected from the 20,000 protein-coding genes. Recent studies have linked abnormally spliced mRNAs with cancerous cells.

In 2012, it was proposed an algorithm for the identification and quantification of polymorphisms of data from RNA-seq when the reference genome is not available without assembling of full transcripts. Although this algorithm identify both approximate tandem repeats, SNPs (single nucleotide polymorphism) and AS, it is only focused on quantifying AS. Due this method, it was possible to realize that annotation of AS events have been underestimated, which 56% of AS identified in the tested dataset were not present in the current notes. However, the algorithm has some limitations. Like most new assemblers based on DBG, (De Bruijn Graphs) the construction of graph requires a very high cost of memory and must be run on a cluster.

In 2016 an article was published which proposes an improvement to an assembler based on DBG. It was removed MPI (message-passing system) and it was implemented Bloom Filter, that is a probabilistic data structure, in the construction of DBG. It was possible run it in a personal computer rather than a cluster. Bloom filter is a probabilistic data structure created by Burton Howard Bloom in 1970, which is used to test whether an element is a member of a set. False positive combinations are possible, but false negative are not, because of that Bloom filter is considered 100% recall rate. That is, it returns 100% of the relevant results.

As the construction of the DBG of this assembler is very similar to that algorithm that indentify and quantify AS, the purpose of this work is the implementation of the Bloom Filter in the identification and quantification AS algorithm, reducing the cost of memory for the creation of DBG, allowing that runs efficiently on a personal computer.

Capturing experimental detail in a paperless environment – Scarab, an Electronic Lab Notebook developed and used at the Structural Genomics Consortium

FERREIRA, L²; MASSIRER, K^{1,2}; GILEADI, O^{2,3}; MARSDEN, B³.

¹*Center for Molecular Biology and Genetic Engineering, University of Campinas, Brazil;* ²*The Strucutural Genomics Consortium, University of Campinas, Brazil;* ³*The Strucutural Genomics Consortium, University Oxford, UK*

The amount of data generated by a biological research team can overcome the ability of that team in preserve the necessary information for the research in an easy way to retrieve it. Developing a data warehouse to store and manipulate the data can help the researchers to improve the quality of the results and also to perform faster researches. In practice, electronic data warehouse allows to store and retrieve all the data relevant to a project, including historical data from past lab members or from collaborators, in a way that is not possible with paper-based notebooks. The development of Scarab by Molsoft L.L.C. in collaboration with the Structural Genomics Consortium – SGC, aims to satisfy this demand for an electronic manner to support research. Scarab incorporates many features centred around structured and unstructured data without losing mining flexibility. Firstly Electronic Lab Notebook – ELN, where the researcher can manipulate electronic pages to embed tables, images, PDF files and also other files used during the research. Those pages also can be used, shared with and modified by researchers who work together on the same project. A unique feature of Scarab is the ability to flexibly mine and export the data, whereby the researcher, without the assistance of an informatics team, is able to build queries to the database and export the result to use the results for another search or attach the result on a desired ELN page or other files. The development of this data warehouse started in 2002 by Molsoft L.L.C. in partnership with SGC located in Oxford University, since then it has been used as a successful case of a bioinformatics tool supporting all activities developed by SGC laboratories around the world. In 2016 the same model was implemented at SGC-Unicamp, proving the modularity of the data warehouse, changing the location and the activities developed at SGC-Unicamp, but using the same data warehouse schema, implementing only minor changes on the database. At SGC-Unicamp we are currently studying the human kinome composed of 500 proteins and we are collecting data from about 10 different sources of experiments. Scarab has proven to present a solid data warehouse solution to support all activities of biological laboratories and the flexibility to adapt to changing science outputs over time.

FUNDING: FAPESP, SGC-Global, Unicamp.

Have you ever wanted to learn about the cladistics origin of operons? Take our TAXI (Taxonomic Innovations)

Ferreira, L. M.¹, Ortega, J.M.²

²*Laboratório de Biodados, Departamento de Bioquímica e de Imunologia, ICB, Universidade Federal de Minas Gerais (UFMG)*

Comparative studies amongst prokaryotic genomes shows genes being shared between them. These genes are not shared only with phylogenetically sister organisms, but are also shared along higher phylogenetic distances. These genes can be translated alone, or grouped into structures denominated operons. Alone or comprising operons, genes can be transferred from one organism to another via horizontal genes transfer. Moreover, genes do not exist in a synthenic orientation within operons in all prokaryotes; there is a clade of origin. In this study were developed a database called TAXI - Taxonomy Innovations, which addresses the innovations along the cladogenesis of bacteria and archaea genomes. This database was built with information from 1752 bacteria and 94 archaea collected from Microbes Online data warehouse. From this data warehouse is was also collected the information of gene clusters, and with the taxonomic distribution we determined the clade of origin of all comprised operons and their genes. For presenting this database we developed a user friendly web interface where the researcher can perform searches for genes or operons and inspect the innovations gained in desired clade. Therefore, knowing the orthologues, their taxonomic/synthetic distribution, ancient *versus* recent genes and operons can be studied. In conclusion, whenever one wants to inspect the epoch of evolution (phylum, class, family, species) in with a gene or operon has first appeared, take our TAXI (Taxonomic Innovations database) available at biodados.icb.ufmg.br/taxi.

NCBI NR Protein Database Clustered by Homology Inference

Aryel Marlus Repula de Oliveira, Roberto Tadeu Raittz

Universidade Federal do Paraná, Setor de Educação Profissional e Tecnológica, Programa de Pós Graduação em Bioinformática

Gene clustering is used to infer proteic features such as homology and function, for example, and it improves the performance of metagenomic analysis, computational searches optimization, etc. Methods for inferring homology from sequences have been an intense topic in research in the past years. The most precise techniques are based in statistical approaches, like Hidden Markov Models allied with sequence similarity comparison. This kind of statistical approach, when used in a huge database like NCBI NR, has two main challenges: firstly, the high computational complexity increases the time spent to perform the clustering and secondly, the training of Markov models without a reference, converges to local maximum of probabilities and do not generate a consistent classification. Due to these challenges, public databases are currently limited either in sensibility or size and the available clustering tools also share these same limitations. We propose a new solution to gene clustering, based on pattern recognition of physical and chemical characteristics of proteins with neural networks, allied with a new and fast algorithm for sequence similarity search that creates a matrix representation of the database sequences, consuming less memory and simplifying the computational processing for comparison and machine learning. Preliminary results demonstrate that the new similarity search algorithm is approximately 4 times faster than the CD-HIT tool, which is used to cluster UniRef, the larger clustered protein database available, producing a similar product than our new algorithm. In this new approach, we create a matrix representation of the NCBI NR protein sequences and start the clustering process considering 50% of similarity; later, we train the neural networks considering, initially, 35 consolidated physical and chemical characteristics and validate the results against curated protein clusters and families. We expect to provide a clustered protein database considering homology inference and also provide a new tool for clustering, using the proposed solution.

Meta-analysis of Japanese Toxicogenomics data: differences between in vivo and in vitro models

Carlos A. O. de Biagi Júnior^{1,2*}, Richa Batra^{5,6}, Jan Baumbach^{3,4}, José L. Rybarczyk Filho^{1,2}

Institute of Biosciences of Botucatu, Univ.Estadual Paulista, ²Institute of Biotechnology, Institute of Biosciences of Botucatu, Univ.Estadual Paulista, ³Department. of Mathematics and Computer Science, University of Southern Denmark, Odense, Denmark,

⁴Computational Systems Biology group, Max Planck Institute for Informatics, Saarbrücken, Germany, ⁵Department of Dermatology and Allergy, Technical University of Munich, Munich, Germany, ⁶Institute of Computational Biology, Helmholtz Zentrum Munich, Munich, Germany

*cbiagijr@ibb.unesp.br

Toxicogenomics is an emerging field to decipher effects of a drug at the molecular level in model systems. One of the main questions is if we can replace the in vivo study by in vitro study. To answer this question we used the data generated by the Japanese Toxicogenomics Project for the Rattus norvegicus liver, with in vivo and in vitro experiments, and Homo sapiens, with only in vitro experiments, treated with 131 drugs (approved by FDA) in different dosages and treatment durations, recorded in a total of 20000 microarray chips. We perform a comparative analysis of the in vivo and in vitro models at modular level using modular map [SEGAL, Eran et al. A module map showing conditional activity of expression modules in cancer. Nature genetics, v. 36, n. 10, p. 1090-1098, 2004.], through a package in R we develop for this methodology, such that each module is a cluster of genes with common gene signature. The analysis made with the comparison between the in vitro Homo sapiens microarray data and REACTOME gene sets, whose data were normalized with the "MAS5" method, yielded a total of 315 clusters. When multiple clusters have similar signatures, we can extract a module from these clusters. This module reflects more clearly the genes that participate in a specific biological process, since it consists of genes whose expression matches the signature of the cluster. Various processes and functions were identified among the clusters obtained previously, including: cell cycle, signal transduction, immune system, metabolism, protein metabolism, gene expression, transport, etc. As a result we obtained a global map showing modules that are induced or repressed in different conditions. The next step of our analysis is to link the modules with clinical conditions.

A database for comparative analysis of paradigms for prospecting contacts in protein-protein interfaces

Martins, P. M. ¹, Mayrink, V. D. ², Silveira, S. A. ³, da Silveira, C. H. ⁴, Lima, L. H. F. ⁵, Melo-Minardi, R. C. ¹

¹ Department of Computer Science, UFMG, Belo Horizonte, Brazil, ² Department of Statistics, UFMG, Belo Horizonte, Brazil, ³ Informatics Department, UFV, Viçosa, Brazil, ⁴ Advanced Campus at Itabira, UNIFEI, Itabira, Brazil, ⁵ Campus at Sete Lagoas, UFSJ, Sete Lagoas, Brazil.

Computing contacts in proteins is important to several types of studies from Bioinformatics to Structural Biology. An accurate computation of contacts is essential to the correctness and reliability of applications involving folding prediction, protein structure prediction, quality assessment of protein structures, network contacts analysis, thermodynamic stability prediction, protein-protein and protein-ligand interactions, docking and so forth. In this work, we built a large database of contacts using about 45,000 PDB files to compare three paradigms for contacts prospection at atomic level: distance-based only, distance and geometric-based (occlusion free) and distance and angulation-based.

The main contribution of this work is a critical evaluation of three different paradigms that can be used to compute contacts between protein atoms. We focused on protein-protein interfaces (multiprotein complex) and analysed four types of contacts, namely hydrogen bonds, aromatic stackings, hydrophobic and ionic (attractive) interactions. We scanned for possible contacts in the range from 0 to 7 Å. Our data showed the importance of a geometric approach to filter out spurious occluded contacts after about 3.5 Å for aromatic stackings, hydrophobic and ionic interactions. For hydrogen bonds the angulation criteria presented more reliable results at every distance in the considered interval. Furthermore, we find a recommended limits of distances and paradigms for each type of interaction considering that approximately 95% of such interactions are in the respective interval.

We provide the database with all computed contacts and the source codes used to populate such database. These resources are available at <http://homepages.dcc.ufmg.br/~pmartins/capri/>.

This work was supported by the Brazilian agencies Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG).

The Polyploid Gene Assembler (PGA)

Leandro Costa do Nascimento^{1,2}, Gonçalo Amarante Guimarães Pereira¹, Marcelo Falsarella Carazzolle¹

1- Laboratório de Genômica e Expressão (LGE) – Departamento de Genética, Evolução e Bioagentes – Instituto de Biologia – Universidade Estadual de Campinas, 2 - Laboratório Central de Tecnologias de Alto Desempenho (LaCTAD) – Universidade Estadual de Campinas

In the last years, hundreds of genomes were sequenced, including species with complex genomes like papaya, panda, orange, etc. The increase of the number of sequenced genomes is directly related with advances in sequencing technologies, which nowadays allows the generation of millions of reads with low costs. In the case of complex genomes, the methodologies for *de novo* assembly remain a bottleneck due to a variety of biological and computational problems. Plant genomes, in particular, have genome size larger than mammals, requiring more elaborated and expansive methodologies for sequencing and high-performance computing to perform the analysis. Moreover, about 80% of the plants have high levels of polyploidy and heterozygosity, which complicate the assembly process for generating drastic variation in the genome sequencing coverage (intronic and intergenic regions have lower coverage than exonic regions). Considering that all genome assembly, including Velvet, SOAPdenovo and Abyss, work with the concept of uniform coverage, i.e., the sequencing coverage remains uniform throughout all regions of the genome, varying only in repetitive regions, *de novo* assembly of plant genomes have resulted in highly fragmented assemblies (millions of contigs lower than 1,000 bp) complicating further analysis, such as gene prediction and annotation. In this context, we present the Polyploid Gene Assembler, a new methodology for reference-assisted sequence assembly focused in genic regions (including UTRs, exons, introns and, in some cases, promoter sequences) using low DNA sequencing coverage (around 3-10x). The pipeline was developed in PERL scripts for running in Linux system that integrates various software for read mapping, *de novo* assembling and scaffolding. In order to solve the assembly problems related to sequencing coverage variations, a *de novo* transcriptome assembly was used because it allows coverage oscillations during De Bruijn graph exploration. Although, PGA was developed for gene assembly from plant genomes, it can be used to any organism that has a closely related species with sequenced genome. PGA has been successfully applied in two complex and very well studied plant genomes: soybean (*Glycine max*) and wheat (*Triticum aestivum*), identifying a total of 99 and 90% of the known genes, respectively. PGA was also used to generate a gene catalogue from *Saccharum officinarum*, an important plant for production of sugar and ethanol, composed by 29,828 transcripts (27,768 genes; mean size of 936 bp), being 27,124 (90.9%) with similarity against the NCBI protein database.

An approach for constructing a database of manually curated contacts in proteins

Pedro M Martins¹, Diego C B Mariano¹, Isabela Pastorini¹, Naiara Pantuza¹,
Marcos F M Silva¹, Raquel C de Melo-Minardi¹

¹ LBS - Laboratory of Bioinformatics and Systems. Department of Computer Science. Federal University of Minas Gerais. Belo Horizonte, Brazil

The comprehension of created patterns through the interaction of atoms and residues, either in proteins or other biomolecules, has been used to solve a range of problems in bioinformatics. For instance: protein folding inferences, structure prediction, functional likeness, structural alignments, thermodynamic stability, protein-ligand and protein-protein interactions, and so on. As a pdb file shows the three-dimensional structure of a protein as x, y and z coordinates of the amino acids, a pair of close residues are considered to be in contact if the distance between their specific atoms is less than a distance threshold. There are a couple of databases where contacts can be found. For instance, Piccolo is a database of structurally-characterized protein-protein interactions described at atomic level. The established contacts are calculated as follows: first, a radial cutoff search is used to identify atoms within 6.05 angstrom. After that, atom pairs are annotated with a specific type of bond depending on the atoms types, distance and the angle between them. However, it is not guaranteed that only the first layer of neighbor atoms are connected by edges and occlusions may occur, making the problem of inferring residue-residue contacts in proteins still unsolved. In this work, we propose an interface to build a manually curated database of protein-protein contacts. The tool counts on a friendly interface that allows the specialists to analyze a contact by visualizing the atoms pairs of residues in separated chains of proteins. Four distinct contact types are considered in the present study: hydrogen bond, hydrophobic interaction, ionic interaction and aromatic stacking. Hydrogen bonds occur between atoms with different electronegativities and have an important role in the enzymatic catalysis. The preference of nonpolar atoms for nonaqueous environments is known as hydrophobic interactions; in globular proteins, hydrophobic effect is important to keep the atoms in an arrangement such that, atoms with higher polarity remain on the surface of the protein and may interact with other molecules, whereas the hydrophobic atoms tend to remain within the protein. Ionic interactions happen between anions (negatively charged atoms) and cations (positively charged atoms). Finally, the aromatic stacks are attractive, noncovalent interactions between aromatic rings and play an important part in the protein folding. The interface also permits a specialist to explain the reasons why (or why not) a contact is truly established and these data may be used afterwards to predict protein-protein contacts through data mining methods.

Supported by:FAPEMIG, CNPq, and CAPES (51/2013 - 23038.004007/2014-82).

Using Data Marts to Select Related Research Articles: A Case Study for the Prioritization of Drug Targets

Marlon Amaro Coelho Teixeira, Kele Teixeira Belloze, Maria Cláudia Cavalcanti,
Floriano Silva-Junior

Acre Federal Institute, CEFET/RJ, Military Institute of Engineering, Oswaldo Cruz Institute,

Protozoan trypanosomiasis are among the ethiological agents of major tropical diseases such as leishmaniasis, Chagas disease, malaria, sleeping sickness and amebiasis. These parasite infections affect the poorest populations of the third world countries with limited access to effective treatments and, therefore, to find novel drugs is of vital importance for them. The research efforts to combat these protozoa grow every day and consequently a large amount of unstructured data has been made available through scientific articles. These articles are accessed in the vast majority of cases by tools that are keyword-based queries, but they are limited and can not meet the needs of researchers. Simple searches performed through these interfaces can return more than a thousand hits. Tools that combine large amounts of data with high performance, enabling users to manipulate and analyse information from different perspectives are more appropriate to deal with this information. However, in the context of a scientific research, these approaches are not quite exploited. The main innovation of this work is to demonstrate that a widely used approach in the analysis of trade data can be applied in analysis of scientific data supporting decision making researcher. Initial experiments were run on a scientific scenario where a corpus of selected papers was annotated using three distinct ontologies with focus on the research of five protozoan organisms: *Entamoeba histolytica*, *Leishmania major*, *Plasmodium falciparum*, *Trypanosoma brucei* and *Trypanosoma cruzi*. Then, the annotation data was extracted, organized and aggregated into a dimensional schema of a demo Data Mart. Finally, based on some simple queries over these data, it was possible to verify that this approach helps the scientist on his/her research, correlating terms and preventing that articles are not accessed. In contrast, using a key-based tool, such query misses many articles and also return many false positives for example, consulting the Gene knockout and knock-out synonymous terms in PubMed, 64017 and 10027 articles are obtained respectively, then if the researcher to use the term knock-out in his query, 53990 articles will no longer be accessed.

ATENA: A decision support system for classifying genetic variants and clinical diagnosis.

Renata Andrade, George Carvalho, Marcel Caraciolo, João Bosco Oliveira

Genomika Diagnósticos

Recent advances in computing and the reductions of costs for next generations sequencing procedures resulted in the generation of a massive amount of data consisting of variants that need to be annotated and interpreted by a specialist. Given the importance of these kind of tests in diagnosing genetic diseases, the American College of Medical Genetics and Genomics (ACMG) published a guideline for interpreting and classifying these variants in order to reduce the variability in the classification and mistakes that could possibly occur. Since the rules in the guideline are complex, they might still lead to inconsistent results when they are assessed manually. In the interest of automating this task, we developed a new tool, ATENA, which evaluates multiple parameters and integrates data from scientific literature in order to satisfy these rules and classify and prioritize variants according to its findings. In order to accomplish this automation, the file originated from the ANNOVAR annotation software is used as an input. The resulting variant is classified as one of 5 possible options: Benign, Likely Benign, Pathogenic, Likely Pathogenic and of Uncertain Significance. When performing the analysis, the specialist can change any of the automatic answers to the evaluated rules in accordance to his personal interpretation, which might lead to a reclassification of the variant automatically performed by the Atena. To evaluate our software, we selected BRCA1 and BRCA2 variants previously classified in the ClinVar database and that had no classification conflicts. Altogether, 325 pathogenic variants and 190 benign ones were chosen. Out of the pathogenic group, 89,5% of the variant were correctly classified while 43,1% of the benign variants fell into the correct class. As a result of the automated classification it was possible to increase productivity and provide grounding for decision making in regards to variant pathogenicity classification.

The importance of an adequate soft-clip based approach on bioinformatics pipelines for multiplex targeted next-generation sequencing

George Carvalho, Renata Andrade, Marcel Caraciolo, João Bosco Oliveira, Rodrigo Bertollo

Genomika Diagnósticos

Advances in high-throughput sequencing have enabled the adoption of sequencing for various applications in research and clinical diagnostics. In addition to lower per-base sequencing costs, one of the crucial factors in reducing per sample sequencing costs is the ability to focus sequencing throughput on specific target regions of interest. One of the main strategies for accomplishing this goal is the use of PCR-based enrichment method by using a few high dimension multiplex PCR reactions (Ampliseq from Life, GeneRead from QIAGEN). The products of PCR enrichment include the primers on both ends. However, these primers are not native to the sample, and need to be removed before variant calling as not to disturb the variant calls from other amplicons that overlap these primers. There are several methods for primer removal, but depending on the strategy selected it might lead to low coverage of reads at the targeted region or missing variants that are located at near the edge of the reads. In some cases, removing the primers at the raw data (FASTQs) can cause misalignments which can lead to a false-positive calls. In this poster, we show that using soft-clipping of the read bases of the primers instead of removing it, it can improve the variant calling sensitivity. We built a custom pipeline for variant calling for amplicon reads using open-source tools such as Cutadapt, BWA, Picard and GATK. For primers base masking we used the tool KATANA, and we compared our results with another pipeline produced by the primers provider. Preliminary tests, conducted with 73 patients, identified 955 variants that compared to the provider's results yielded 85.23% true-positive, 5.13% false-positive and 9.63% false-negative rates. Among the false-positives, approximately 1% of the variants were true and all false-negatives were the results of bad trimming on the provider's part. We outperformed the results of the provider pipeline reducing the number of false-positives and false-negatives due to incorrect primer masking and missing low coverage variant calls. In this poster we would like to share with the audience the lessons learned during the development and present the best practices and strategies to work with amplicon reads in variant calling pipelines.

visGReMLIN: An interactive strategy to visualize common substructures in protein-ligand interaction

Vagner S. Ribeiro¹, Charles A. Santana¹, Fabio R. Cerqueira¹,
Alexandre V. Fassio², Carlos H. da Silveira³, Raquel C. de Melo-Minardi²,
Sabrina de A. Silveira¹

¹*Informatics Department, Universidade Federal de Viçosa*

²*Computer Science Department, Universidade Federal de Minas Gerais*

³*Universidade Federal de Itajubá*

Interactions between proteins and ligands play an important role in biological processes of living systems. The comprehension of protein-ligand molecular recognition is an important step to ligand prediction, target identification and drug design, among others. Currently, we have a visual interactive interface to explore protein-ligand interactions and their conserved substructures for a set of similar proteins, which we named visGReMLIN, that allows to visualize protein-ligand interaction patterns computed by GReMLIN (GRaph Mining strategy to infer protein-Ligand INteraction patterns). This tool shows patterns in protein-ligand interaction for two test datasets: (i) CDK2 which comprehends 73 entries from Protein Data Bank (PDB) with identical sequences coupled with different ligands. CDK2 has an important role in cell cycle regulation; (ii) Ricin, 29 PDB entries, which share sequence identity greater than or equal 50% with ricin template 2AAI chain A. Ricin is a notorious protein that acts as a potent toxin. GReMLIN uses a strategy based on frequent subgraph mining, that is able to perceive structural arrangements relevant for protein-ligand interaction. In this abstract, we propose a generic version of visGReMLIN, in a way that it will enable biologists, biochemists and anyone who has interest in protein-ligand interactions to search and visualize patterns in compound structures from their own dataset of interest or from structures directly downloaded from PDB. With the PDB entries in hands, we will execute all steps from GreMLIN strategy to compute protein-ligand interaction patterns. Meanwhile, we give users a job number, to return to our website and explore the results when they are ready, as patterns computation can take some minutes.

Financial support: CAPES, CNPq, FAPEMIG

An automated method for the identification of Dengue, Zika, Yellow Fever and Chikungunya virus species and genotypes

Luiz Carlos Júnior Alcântara¹, Nuno R. Faria², Marta Giovanetti¹, Vagner Fonseca¹,
Maria Inés Restovic¹, Murilo Freire¹ & Túlio de Oliveira³

¹*Oswaldo Cruz Foundation (FIOCRUZ), Salvador, Bahia, Brazil.* ²*Oxford University, UK.* ³*Africa Centre, University of KwaZulu-Natal, Durban, South Africa.*

In recent years, an increasing number of outbreaks of Dengue virus (DENV), Zika virus (ZIKV) and Chikungunya virus (CHIKV) have been reported in Asia and the Americas, while the Yellow Fever virus (YFV) continues endemic in Africa. The geographical distribution of ZIKV has expanded significantly reported now in at least 41 countries. Since these arboviruses share many clinical symptoms, such as febrile illness with rash, myalgia, or arthralgia, and current serological tests lack the power to discriminate between ZIKV and other flaviviruses, such as DENV and YFV, genetic testing during acute infection has become the standard method to identify the cause of infection. To facilitate diagnosis and the development of prevention and treatment strategies that efficiently target the diversity of these viruses, we developed a rapid high-throughput-genotyping system. The method involves the alignment of a query sequence with a carefully selected set of predefined reference strains, followed by phylogenetic analysis of multiple overlapping segments of the alignment using a sliding window. Each segment of the query sequence is assigned the genotype and sub-genotype of the reference strain with the highest bootstrap (>70%) and bootscanning (>90%) scores. The new Arbovirus-Genotyping Tools provide accurate classification of these arboviruses and are currently being assessed for their diagnostic utility. In conclusion, our new computational method allows the high-throughput classification of DENV, ZIKV, YFV and CHIKV species and genotypes in seconds. Species can be classified using short reads from any NGS platform, such as metagenomics Illumina's RNA-seq, and genotypes can be classified most confidently when using envelope gene or complete genome sequences. The framework's is freely available online from a dedicated server (<http://www.bioafrica.net/software.php>).

Area: Software Development and Databases

Dugong: a Docker image, inspired on Ubuntu Linux, designed to enhance reproducibility and replicability during computational analyses of biological data

Fabiano Bezerra Menegidio , Luiz R. Nunes

Núcleo Integrado de Biotecnologia, Universidade de Mogi das Cruzes, Brasil, Centro de Ciências Naturais e Humanas, Universidade Federal do ABC, Santo André, Brasil

The increasing use of computational methods for the analysis of biological data and the constant expansion verified in the fields of Bioinformatics and Computational Biology have revolutionized the study of Biology during the past few decades. However, grasping the complex nature of some softwares employed for such analyses and adapting to the rapid changes observed in computational ecosystems has become a major challenge for biologists, hampering the full use of such resources by the scientific community. This problem becomes more serious due to the fact that many computational methods often rely on pipelines composed by multiple analytical steps, involving different scripts, softwares and/or algorithms with unique requirements and/or dependencies. As a result, utilization of computational resources during bioinformatics analyses has become increasingly heterogeneous across laboratories, compromising reproducibility and replicability of results obtained from a given experiment or dataset. Although the Bioinformatics community has heavily relied on the production of Open Source softwares as a way to minimize these problems, mandatory installation of libraries for the proper functioning of scripts, lack of proper documentation and incompatibility with different operating systems and/or hardware still represent major obstacles to ensure replicability and reproducibility of data analysis in different computational environments. Fortunately, emergence of the Docker project is providing a promising new strategy to tackle these problems, by allowing the configuration of a complete computing environment, in which all libraries, codes and additional data required for a particular application may be implemented in a single container, which can be consistently exchanged and launched in different platforms, regardless the specificities of their hardware and/or operating systems. Thus, to explore and demonstrate the usefulness of Docker-based systems as a strategy to enhance replicability and reproducibility of bioinformatics analyses in multiple computing environments, we developed the application Dugong, a Docker image based on Ubuntu 15.10, specifically designed for the analysis of large-scale biological data. Using a graphic interface generated by Xfce4, Dugong provides the managers Linuxbrew (Homebrew Science) and Conda (Bioconda), which allow distribution and installation of over 3000 bioinformatics-related packages and libraries, with automated installation of their respective dependencies. Simulations performed in virtual machines demonstrate that Dugong allows effective creation of reusable containers for different bioinformatics analyses in a uniform computational environment, allowing acquisition of consistent and reproducible results by the scientific community, thus assisting in the development of Open Science projects.

CEMiTool: Co-Expression Modules Identification Tool

Pedro S. T. Russo^{1,2}; Gustavo Rodrigues-Ferreira¹; César A. Prada-Medina¹; Matheus C. Bürger^{1,2}; Lucas Esteves Cardozo¹, Luciane Schons-Fonseca³; Thiago D. C. Hirata¹; Gonzalo Sepúlveda-Hermosilla⁴, Vinícius Maracajá-Coutinho⁴; Helder I. Nakaya^{1,2}

1 - School of Pharmaceutical Sciences – University of São Paulo, São Paulo, SP, Brasil; 2 - Graduate Program in Bioinformatics - Institute of Mathematics and Statistics - University of São Paulo, SP, Brasil; 3 - Department of Biology - Massachussets Institute of Technology, Cambridge, MA, USA; 4 - University of Talca, Talca, Chile

Systems Biology approaches provide a holistic view of biological processes, integrating the many different molecular cell components through highly complex networks. By analyzing these networks, it is possible to create mathematical and computational models that help understand the mechanisms triggered by different types of stimuli or perturbations. Also, predictive modelling can be applied to reveal gene signatures associated with systemic responses, such as disease outcome or vaccine-induced immunity. However, in addition to the stochasticity of biological processes and the noise associated with high-throughput technologies, gene networks are very dynamic and highly sensitive to changes in conditions and experimental settings. Therefore, given the inherent modularity of biological systems, in this project we developed an easy-to-use web-based application called CEMiTool, which aims to identify the underlying modules in gene co-expression networks and analyze their changes in response to different types of stimuli and perturbations. CEMiTool's code is written completely in R programming language. In order to test our tool, we ran CEMiTool on Juvenile Idiopathic Arthritis, inflammatory bowel diseases (ulcerative colitis and Crohn's disease) and sepsis studies. Our analyses revealed several genes coexpressed in the different inflammatory diseases. Modules included genes associated with inflammatory and anti-viral pathways, such as type I interferons, NOD1/2 signaling pathways and TNFR1-induced NFκB signaling pathways. We found that some of the gene modules were unique to a disease whereas some modules were shared by different diseases. We also associate the activity of the modules with the inflammatory score of patients. This analysis provided novel insights to disease mechanisms.

Cancer immunology of Cutaneous Melanoma: A Systems Biology Approach

Munoz, Mindy; Nakaya, Helder.

*Computational Systems Biology Laboratory – Faculdade de Ciências Farmacêuticas,
Universidade de São Paulo.*

Cutaneous melanoma is a melanocyte skin cancer and it is one of the most aggressive tumors in humans. It causes a great number of deaths worldwide, and in Brazil approximately 1,300 melanoma patients die each year. The Cancer Genome Atlas (TCGA) database contains genomics, epigenomics and transcriptomics data from 470 samples of skin cutaneous melanoma (SKCM). Few studies have applied systems biology approaches to investigate melanoma progression. However, they failed in integrating several layers of “omics” data in order to elucidate the mechanisms by which melanoma cells become resistant to the immune system. We propose here to perform an integrative omics analysis with the SKCM data available in TCGA. For that, we will utilize established network models coupled with hub detection algorithms. We constructed gene network integrating human databases from IntAct, BioGRID and HPRD. And the interaction data was limited of validated protein-protein interactions with experimental data from our lab and curated from scientific literature. This integrated data was represented as an undirected network, where each node represents a human gene and each edge represents a pair of genes, like binary interaction in the human interactome. If exists a physical interaction between genes an edge is connected. Network centrality analysis was carried out by means of calculating measures of centrality for each gene in the interactome. We take a assigned immunological score by relevance of genes in tumor immunity and melanoma and analyze transcriptomic, genetic and epigenetic profiling, with phenotypic characteristics throw clinical data of patients, joined to the identification of hub genes can help us to unravel the role of immune system in SKCM progression, focusing in personalized medicine.

Global coexpression analysis of human protein-coding genes

Katia de Paiva Lopes^{1,2}, Francisco José Campos-Laborie², Ricardo Assunção Vialle¹, José Miguel Ortega¹, Javier De Las Rivas²

¹Departamento de Bioquímica e Imunologia, Instituto de Ciências Biológicas (ICB), Universidade Federal de Minas Gerais (UFMG), Brasil.

²Bioinformatics and Functional Genomics Group, Cancer Research Center (CiC-IBMCC, CSIC/USAL/IBSAL), Consejo Superior de Investigaciones Científicas (CSIC), Salamanca, Spain

Advances in high throughput sequencing technologies have introduced a new alternative to transcriptome analysis, namely RNA-Seq. Studies using this method have already altered our view of the extent and complexity of eukaryotic transcriptome. However, until recently, little has been reported about the determinants of human cell identity, particularly from the joint perspective of gene evolution and expression. In view of these, our work presents a combined analysis of human transcriptome data: 1) An evolutionary analysis using a RNA-Seq dataset of 116 samples from 32 tissues (E-MTAB-2836) using a database of orthologous proteins (OMA); and 2) a relational context of the human protein-coding genes based on a robust coexpression network analysis. Therefore, we present a complex network -like a galaxy- that includes 1,691 protein nodes related with 19,615 interactions. This network corresponds to a subset of the coexpression network, which includes 2,298 proteins and 20,005 interactions. The coexpression dataset was built calculating the pair-wise Spearman correlation coefficient (r) of all the genes along the 116 samples and only selecting, as positive gene-pairs, the ones that had a correlation coefficient ≥ 0.85 . A cross-validation of these correlation values was also applied by a random selection of two sample replicates from each tissue (i.e. a total of $32 \times 2 = 64$ samples) and recalculating again the Spearman correlation for these random subsets of the data. This sampling was done 100 times, annotating for each gene-pair the number of times that its r coefficient was ≥ 0.85 . Only the gene-pairs validated 100 time in this sampling were selected. The analysis of the network done with MCODE revealed the existence of 11 major subnetworks -considered as major constellations in the galaxy of nodes- that had a clear enrichment in certain groups or modules of highly coexpressed proteins showing a tendency to include proteins of the same evolutionary age. Finally, the study of the pair-wise correlation of the gene expression profiles along tissues allowed building human gene coexpression networks and find modules with functional and biological meaning where we did map the age of the genes and demonstrate the existence of tighter links between age-related proteins.

Supported by: Capes, FAPEMIG, CNPq.

Meta-dimensional analysis in gene network inference and gene prioritization associated to complex diseases

Carlos Eduardo Marchi, David Corrêa Martins Junior, Fábio Marchi

Federal University of ABC, A.C.Camargo Cancer Center

Gene prioritization approaches are rich tools for personalized medicine and system biology, helping to reveal genes involved in an abnormal condition and being a decisive factor in the diagnosis, the advancement of therapeutic practices and the development of less aggressive and more effective drugs. In this context, we are developing a computational system for gene prioritization which integrates a number of biological data sources in order to reveal genes associated to complex diseases. To pursue this goal, this system implements integrative methodologies based on meta-dimensional data analysis, such as concatenation and transformation to integrate gene expression, miRNA expression and DNA methylation data under case and control conditions. Furthermore, it has involved inference and analysis of complex networks in order to rank candidate genes to be associate to a given disease. The data preprocessing consists of cleaning and removal of outliers. After that, a technique of differential analysis involving the clinical data of the patients are applying to obtain the most promising genes. Next, several gene correlation networks are created for each data type and also by the combination of them, such as mRNA-miRNA, mRNA-DNA methylation and mRNA-miRNA-DNA methylation. Pearson and Spearman correlation methods can be applied in the networks inference process. Finally, the topological analysis is executed to rank genes, including several algorithms and criteria based on complex network theory, such as degree and betweenness centrality; Page Rank, Hyperlink-Induced Topic Search and Katz centrality. This topological analysis results in a scored gene list for each network, which can be analyzed together to create a final list of the top ranked genes. Our method is unique integrative approach that involves the biological data sources aforementioned and that focused on clinical data of patients, enabling future analysis to develop new drugs and treatments for groups of patients. We expect that the evolution of this system could be an important step toward personalized medicine useful to improve the understanding of a myriad of complex diseases, including cancers, neurodegenerative and neurodevelopment disorders.

Funding support: UFABC, CAPES

Logical Modeling of Cellular Senescence Induced by DNA Damage and TGF β signaling

Veronica Venturini Rossato, José Carlos Merino Mombach

Departamento de Física - Universidade Federal de Santa Maria, RS, Brasil

Upon DNA damage or by external stimuli (like TGF β) a cell can activate cell cycle checkpoints (mechanisms that arrest the cell cycle to ensure that the required steps are satisfied prior to a phase transition) which promote DNA repair, senescence or apoptosis. Senescent cells release factors that induce permanent senescence as a bystander effect in the cellular environment. These factors are known as SMS (Senescence-messaging secretome), and among them, TGF β is an important component involved in the 'bystander senescence'. In this work we present an expanded version of the logical model proposed by Mombach *et al.* in 2015 by including an input representing the influence of TGF β . The model contemplates the crosstalk between TGF β and DNA damage pathways that are important in inducing cellular senescence. In a logical model the variables representing proteins are discrete and the interactions among them are represented by logical operators (And, Or and Not). The model inputs are: TGF β level of stimulation, repairable or irreparable SSB (DNA single Strand Breaks) and DSB (DNA double Strand Breaks). The outputs of the model are the following phenotypes: senescence, apoptosis, cell cycle arrest and proliferation. Mutations representing gain of function or loss of function of proteins were produced using the tool GINsim 2.9 and the simulations have demonstrated consistency with the experimental literature on cell growth phenotypes obtained from mutant cells. We also observe the crosstalk effect enhancing senescence and apoptosis upon the combined stimulation of TGF β and irreparable DNA damage.

Support: CNPq and Capes

Saccharomyces cerevisiae Protein-Protein Interaction Network Reconstruction to Study Ethanol Tolerance

Ivan R. Wolf, Lauana Fogaça, Leonardo Nazário de Moraes, Rejane M. T.Grotto,
Rafael P. Simões, Guilherme Valente

*Department of Bioprocess and Biotechnology. São Paulo State University (UNESP),
Botucatu*

The interest to develop biofuels such as bioethanol have been increasing. The most common process to produce bioethanol is the first generation technology, in which the most widely used organism is *Saccharomyces cerevisiae*. However, the high concentration of ethanol produces toxicity to *S. cerevisiae*, which is the main limiting factor to produce this fuel. Despite lots of effort to understand this phenomenon, the ethanol tolerance on the systemic view point is poorly understood. In the present study, the highest ethanol tolerance was experimentally determined for five yeast strains (S288c, BY4741, BY4742, SEY6210 and X2180-1A), in which unsupervised learning was applied over the experimental data matrix to classify the strains as highest tolerant or lowest tolerant. For the reconstruction of protein-protein interactions (PPIs) networks, all protein sequences for those strains were obtained from the *Saccharomyces* Genome Database and all protein pairs were submitted to UNISPPI PPI prediction, in which each protein pair was defined as an edge of the network. Three scores were calculated for each edge: (I) probability of protein interaction from UNISPPI; (II) the percentage of shared sub-cellular locations in YeastGFP database; (III) the percentage of shared sub-cellular locations in ComPPI database. In order to filter out the edges, they were split into four datasets: (A) edges with scores I, II and III; (B) interactions with scores I and II; (C) edges with scores I and III; (D) edges only with score I. The Principal Component Analysis (PCA) was applied on each dataset, independently, and edges inside the limits established from first to fourth quartile over PCA coordinates were selected. The dataset D consists only of UNISPPI score, and in this case only edges with score ≥ 0.724 were selected. The procedure was efficient to reconstruct the networks since the degree distribution follow a power law, fitting the Barabási-Albert model; and they also have a scale-free topology. The average degree and average betweenness centrality were high for all nets, which express the presence of hubs. Moreover, the networks are assortative, which means that the nodes with high degree are preferentially connected on each other, following the principle of “rich-get-richer” dynamics. Those global topological characteristics fit with the expectations of a real biological network. Furthermore, a clustering analysis over all nodes degree for all networks showed that the degree is a property closely related to the highest ethanol tolerance on yeast (clusters obey the percentages of ethanol tolerance).

Funding Support: BIOEN FAPESP and PROPe UNESP.

A logical model for the bimodal p53 switch in cell-fate control

Maria Vitória Cavalheiro Issler, José Carlos Merino Mombach

Departamento de Física – Universidade Federal de Santa Maria, RS, Brasil

The p53 pathway is activated in response to DNA damage, leading to different cell fate decisions, as cell cycle arrest and the possibility of DNA repair, senescence or apoptosis. Recent experimental works have suggested that is not just the strength of damage that controls cell fate, but rather the dynamics of p53. Chen et. Al (2013) have studied U-2 OS cells under DNA damage generated by etoposide which is a chemotherapeutic compound that induces DNA damage and is also known to activate the kinase p38MAPK. They identified a bimodal switch of p53 dynamics due to MDM2 upregulation that is important for cell-fate control. Low levels of damage (low concentration of etoposide) primarily induce p53 pulses and the cell undergo cell cycle arrest, whereas high level of damage (high concentration of etoposide) might induce a p53 monotonic increase and the cell eventually enters apoptosis. Given this background, in this work we propose a logical model for the switch behavior in p53 dynamics that contemplates the upstream kinases induced by DNA damage that regulate p53 as ATM/ATR and other important p53 regulators like p38MAPK, Wip1 and MDM2 (nuclear and cytoplasmic). In a logical model the proteins have discrete state values and the interactions are represented by the logical operators AND, OR and NOT. The input of the model is the level of DNA damage. Simulations of the model were generated using the tool GINsim 2.9.4. We found that for high DNA damage the model presents bistable dynamics, where one of the stable states presents p53 in its highest activation level corresponding to apoptosis, while the other state has p53 in its lower activation level corresponding possibly to senescence. Both stable states present inhibition of nuclear MDM2 which is observed experimentally. For the intermediate level of damage, the model presents a terminal cycle, where p53 oscillates between levels 0 and 1, never reaching level 2, and p21 and MDM2 also oscillate. In the model, knockdown of nuclear MDM2 leads to stable states and completely abrogates the terminal cycle, independently of the DNA damage level, showing the crucial influence of MDM2 on p53 dynamics.

Support: CNPq

Pattern Recognition in genomic sequences: A case of study using complex networks

Isaque Katahira^{1,2}, Fabrício Martins Lopes¹

¹*Universidade Tecnológica Federal do Paraná – Campus Cornélio Procópio
Departamento de Computação*

Programa de Pós – Graduação em Bioinformática – PPGBIOINFO

²*Centro Paula Souza – Escola Técnica Estadual Professor Mário Antônio Verza*

The integration of computing with other knowledge fields is becoming increasingly important for the research development in various fields of science. Notably, Bioinformatics is an example of this fact, since it is inherently multidisciplinary by combining knowledge of biology, statistics and computer science among other relevant sciences. In this context, the simultaneous extraction of molecular data from thousands of genes is a major breakthrough in the area, but also a challenge, as it generates a huge volume of biological data to be analyzed. In particular, it is essential to develop new methods and techniques in order to understand the influence of gene expression in the functional state of an organism. In this context, a possible way that can contribute is to develop techniques able to reduce the amount of data without loss of information contained therein. In this sense, the feature extraction techniques can be applied, which can be used to represent the data through its most relevant properties without using the entire data set. Thus, it is proposed to develop an approach based on mapping of genome sequences in their complex networks representation. From these networks can be taken measures to specify them into feature vectors which may be used to summarize the data collected, in order to quantify the topological similarity between the generated networks. It is justified, therefore, the use of complex networks for its approach to the real networks with respect to nonlinearity, where structural arrangements of its nodes can be crucial for understanding the interactions, functions and its hole structure. The observation of the extracted characteristics via nodes, edges, and other arrangements may provide the identification of patterns contained in these networks. Therefore, it is expected that the network measures leads to distinguish different classes of genomic sequences, turns possible to verify mutations, help in constructing phylogenetic trees, and other analyzes. Thus, there is the potential to observe different patterns contained in biological structures and propose large-scale rating of applications in the context of systems biology, that is, considering the modeling of an organism as a whole.

Influence of a high-fat diet in the cerebellar tissue of Cockayne Syndrome mice

Gabriel Baldissera, Kendi Nishino Miyamoto, Diego Bonatto

Computational and Molecular Laboratory, Biotechnology Center, Rio Grande do Sul

Federal University

Cockayne syndrome (CS) is an autosomal recessive progeroid neurodegenerative disease. Cerebellum is the most affected brain's region with observed irregular myelination patterns. The syndrome is caused by mutations in both CSA and CSB genes, involved in DNA repair mechanism. Both mutations changes the activities of interacting proteins, like DNA repair signaling protein PARP. It was observed that PARP is overactivated in CS, resulting in ATP depletion and cell death. Recent data demonstrates that a high-fat diet (HFD) can rescue the CS symptoms and consequently improve cerebellar functions. However, the mechanisms of HFD-inducing cerebellar protection are unknown. Therefore, this work aims to understand the major genes that are modulated in cerebellum of a CS-model mouse fed with a HFD. Microarray data (GSE62194) was downloaded from Gene Expression Omnibus (GEO) into "R" statistical environment, being quality-assessed and statistically analyzed by *arrayQualityMetrics* and *limma* packages, respectively. The gene expression comparison between HFD fed and standard diet fed CS-model mice generated a list of differentially expressed genes (DEGs) that was used as an input in the metasearch website STRING in order to generate an interatomic network. Additionally, the lipids found in HFD were utilized as an input in the website STITCH to draw chemical interaction networks. These networks were merged in a main network using the software Cytoscape, in which were performed clustering, centralities and gene ontologies (GO) analyses. The clustering and GO analysis indicated two main clusters associated to lipid transport and metabolism, as well as cell death regulation. It was also observed in the main network the presence of overexpressed genes associated to lipid metabolism and Krebs cycle, suggesting the positive regulation of lipid catabolism in CS. In addition, LINGO1, which is associated to myelination and oligodendrocyte differentiation, was found differentially expressed in HFD fed mice. The data supports the hypothesis that lipid catabolism could be necessary for providing a positive energetic balance for CS-affected neuron cells.

Funding support: FAPERGS

ANOVA-like method for differential correlation of multiple networks analysis of biological data

Vinícius Jardim Carvalho 1,2, Suzana de Siqueira Santos 3, Adriana Grandis 4,
Amanda Pereira de Souza 5, André Fujita 3, Marcos Silveira Buckeridge 2

1 Bioinformatics Graduate Program, University of São Paulo, São Paulo, SP, Brazil, **2**
Department of Botany, Institute of Biosciences, University of São Paulo, São Paulo,
Brazil, **3** Department of Computer Science, Institute of Mathematics and Statistics,
University of São Paulo, São Paulo, Brazil, **4** Institute for Genomic Biology, University of
Illinois at Urbana-Champaign, Urbana, IL, 61801, USA, **5** National Institute of Amazon
Research, Manaus, Brazil

Identify if metabolites or genes expression patterns range over an experiment in response to environmental conditions is a major task in Bioinformatics. Therefore statistical tests such as t-test and ANOVA, are used to identify which variables significantly range between two or more biological conditions. However, those tests do not take into account the information about the relationships among variables. To overcome that limitation, several multivariate methods were developed such as PCA, cluster analysis, multiple regression, and network analysis such as the CoGA software. Network analysis allows us to address the connectivity between studied variables. The CoGA software performs differential network analysis between two graphs (one for each biological condition) based on network topological characteristics, such as centralities, clustering coefficient and spectral distribution. Despite the important role of CoGA, plant physiology experiments often compare more than two biological states. In order to fill this gap, we aim to implement a generalization of the CoGA method for two or more graphs, which can be very useful when we compare many biological states. The ANOVA-like test for several graphs performs based on the mean of Kullback-Leibler divergences between their spectral distributions (distributions of the eigenvalues of the graph adjacency matrices) and the average distribution of all graphs. The pvalue of statistical test for the divergence is made through permutation of the labels. In order to evaluate if the spectral distribution test controls the false positive rate and to measure its statistical power we performed simulation experiments with biological data. Thus the proposed method can bring evidences of differences in the network structure among two or more biological conditions. The application of our method to two sets of data on plant physiology and biochemistry obtained from a C4 and a C3 plant under different experimental conditions revealed that the method is reliable for robust statistical comparisons among networks either within 24h or along several weeks. We expect that this method will be useful for plant (and animal) physiologists, providing means to analyze large data sets that integrate molecular, biochemical and physiological data so that whole organisms could be scientifically compared under different environmental conditions.

Using Systems Biology to Understand Immunosenescence

Fernando Marcon Passos¹, Pedro de Sá Tavares Russo¹, Matheus Carvalho Bürger¹, Thiago Dominguez Crespo Hirata², Dr. Helder Takashi Imoto Nakaya²

¹*Bioinformatics Graduate Program – Institute of Mathematics and Statistics - University of São Paulo - São Paulo, Brazil,* ²*Physiopathology and Toxicology Graduate Program – School of Pharmaceutical Sciences – University of São Paulo - São Paulo, Brazil*

The remodeling of the immune system that comes with age, known as immunosenescence, contributes to an increased susceptibility in elderly to infectious diseases, cancer, autoimmunity and decreased vaccines response. This remodeling is a complex and multifactorial process and, until now, there is little understanding of the molecular mechanisms involved. Several studies tried to understand and identify which genes and signaling pathways are involved in the ageing of our immune system. However, none has yet done a comprehensive analysis of a large amount of transcriptomic data of healthy subjects in a wide age spectrum. In this project we aim to create a predictive model for the biological age of the immune system using machine learning methods. For that we will perform a meta-analyses of microarray transcriptomic data available in the GEO public repository. We selected 29 studies containing 435 blood samples that had subject's age information. First, we will identify genes that are differentially expressed between age groups, through the statistical method LIMMA. With such genes we can discover the gene signatures that are related with immunosenescence throughout a pathway enrichment analysis. In this step, we will also perform a coexpression analysis to build gene networks related to immunosenescence. This will be done using the CEMiTool, a tool developed in our laboratory that allow us to identify gene modules and sub-modules associated with a particular phenotype. The next step is to create a predictive model of the biological age of the immune system. We will use dimensionality reduction and feature selection algorithms, like PCA and the FSelector package, to select genes that optimize the predictive power of the model. Then, we will use various algorithms of machine learning, such as Support Vector Machine and Neural Networks, to create age group classification and age regression models. These models will then be validated with blood samples from children and elderly, which will be provided by the Liverpool School of Tropical Medicine. Once the model has been validated, genes used in machine learning algorithm as well as the regulation profile and co-expression of gene networks discovered in the meta-analysis will be used to understand the mechanisms of activation and deactivation of the genes they are related to immunosenescence.

Construction of metabolic map in lead poisoning

Souza, I.D.¹, Andrade, A.S.¹, Dalmolin, R.J.S.¹

¹*Programa de Pós-Graduação em Bioinformática, Universidade Federal do Rio Grande do Norte*

Since antiquity, lead was used by humans as a tool for daily routine, due to its unique properties, like malleability, ductility, corrosion resistance, low melting point and low electrical conductivity. Nowadays, lead is one of the most studied heavy metals worldwide due to its relation with occupational exposure on industry workers or people who manipulate its compounds. Many effects of lead poisoning have already been reported on literature, showing a compromising of whole body health, with symptoms related to cardiovascular, immune, bone, reproductive, hematological, renal and gastrointestinal systems, even the most studied ones are related with neurological system. Although there is evidence of how it affects homeostasis at the cellular level, the description of metabolic pathways affected in lead poisoning is not fully established. To clarify the effects of lead poisoning, the aim of this study is to build a metabolic map of the cellular and biochemical pathways altered by the presence of lead, and to analyze which proteins and cellular components this heavy metal has the ability to bind and interfere with its normal function. For this purpose, we made a textmining search in literature in order to obtain information about lead interactions with biomolecules. On our preliminary results, we have already found a total of 22 proteins which can directly interact with lead, with many other proteins that can be indirectly affected by this metal. In order to validate our previous observations on literature, we analyzed some data of lead intoxication of experimental models from public repositories. The metabolic map using the pathways affected by lead will then be built by the comparison of these two sources.

RTNsurvival: An R package for making survival analyses and plots from transcriptional networks

Clarice S. Groeneveld, Vinícius S. Chagas, Kelin G. Oliveira, Mauro A. A. Castro

Bioinformatics and Systems Biology Laboratory, Federal University of Paraná (UFPR)

Transcriptional networks are important objects of study to better understand the complexity of interactions between genes such as how some key transcriptional factors act in a specific phenotype. Several methods can be used to reconstruct and understand these networks, one is the RTN package, an R tool to build transcriptional network of regulons centered around any regulatory element, such as transcription factors and miRNAs. These regulons can be used to interpret a wide variety of data, including clinical information. However, there is currently no simple tool to integrate survival analysis with reconstructed transcriptional networks ("TNI" objects) from RTN. Here we show the RTNsurvival, which is an R package for making survival analyses and plots from "TNI" object. The package takes transcriptional networks generated by RTN and integrates it with survival data from the same cohort, the methods consist in a Kaplan-Meier curve and a Proportional Hazards Regression model (Cox) analysis, which are regulon-specific, and their respective plots. The package was used to analyse data from a cohort of breast cancer (BRCA) patients from the METABRIC database with a previously computed 'TNI' with transcription factor regulons. It confirmed that the regulon centered around of the estrogen receptor gene (ESR1) is predictive of survival in BRCA patients, with patients with active regulons having higher survival and patients with a repressed regulon having lower survival. RTNsurvival will be submitted to Bioconductor and integrate the RTN-family of R packages, with the intent of facilitating further survival analyses of transcriptional networks provided by RTN package.

Interactome analysis of FGFR2 – a potential therapeutic target in breast cancer.

Kelin G. DeOliveira, Mauro A. A. Castro.

Universidade Federal do Paraná (UFPR), Setor de Educação Profissional e Tecnológica, Programa de Pós Graduação em Bioinformática, Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq)

Breast cancer (BC) is a leading cause of cancer death among women and results from multivariate factors of environmental and genetic origin. One of most consistent genetic risk factors for breast cancer is the single nucleotide polymorphism (SNP) rs2981582 that sits inside the intron 2 of the Fibroblast Growth Factor Receptor 2 (FGFR2) gene. The encoded protein, fgfr2, is a tyrosine kinase receptor that acts as a housekeeping molecule at the mammary tissue – amongst other tissues – and it is implicated in several cellular processes, such as cell cycle control, proliferation, migration, tissue repair and tumorigenesis. It has been demonstrated that the risk SNP rs2981582 is able to modulate the expression of the FGFR2 gene. However, the downstream effects in breast cancer at protein level are yet to be clarified. Here we used a data mining approach to reconstruct the fgfr2 interactome from a wide range of databases (STRING, HRPD, IntAct, BIOGRID, DIP, I2D and MINT). The resulting curated protein-protein interaction (PPI) network comprises the most significant proteins that establish interactions with the fgfr2 protein in several tissues and in non-tumor conditions. Along with previous studies, the fgfr2 PPI network might provide a rich groundwork to address systemic questions on how fgfr2 might affect signaling pathways related with breast cancer. Once we have concluded the curated PPI network, proper trimming will then be performed in order to map a mammary tissue specific fgfr2 PPI network, which will be validated through knockdown experimental data. We anticipate that this tissue-specific PPI network will allow us to exhaustively explore other relevant breast cancer risk SNPs in the context of fgfr2 signalling events.

Understanding transcriptional strategy for Inositol pathway in soybean root dehydration stress tolerance

Bezerra-Neto, João Pacifico^{1,3}; Ferreira-Neto, José Ribamar Costa^{1,2}; Kido, Ederson Akio²; Silva, Manassés Daniel²; Benko-Iseppon, Ana Maria¹; Santos, Mauro Guida³.

¹Laboratory of Plant Genetics and Biotechnology, Genetics Department, Universidade Federal de Pernambuco, Recife, PE, Brazil; ²Laboratory of Molecular Genetics, Genetics Department, Universidade Federal de Pernambuco, Recife, PE, Brazil; ³Laboratory of Plant Physiology, Department of Botany, Universidade Federal de Pernambuco, Recife, PE, Brazil

Throughout the evolutionary process, plants have developed a range of molecular mechanisms to withstand water deficit conditions. Inositol (Ins) is a polyalcohol, a cyclic carbohydrate with six hydroxyl groups one on each of the ring carbons, which has multiple effects on plant metabolism, acting since the production of secondary messengers to the synthesis of osmolytes/antioxidants. Structural genomics and global transcriptome analyses associated to inositol are not available for plants, except regarding genes involved in the metabolism of raffinose in maize. The present work aimed to identify loci linked to Ins metabolism in the soybean genome, besides globally cataloging and analyzing the related transcripts orchestration that acts increasing the pathways efficiency. Using the HT-SuperSAGE experimental data, tags (26 bp) were analyzed to determine unique tags (unitags) and those differentially expressed ($p < 0.05$), while singlets (tags only sequenced once) were excluded from the evaluation. Unitags were annotated by BLASTn against transcripts sequences from Phytozome database (<http://www.phytozome.org/>), anchored in the Phytozome genome, and when the equivalent annotated sequence was an enzyme, the available EC number was recovered. The in silico biochemical pathway analysis involved the Kyoto Encyclopedia of Genes and Genomes (KEGG) Pathway database (<http://www.genome.jp/kegg/pathway.html>) applying for the EC numbers in the KEGG Mapper – Search & Color Pathway tool, trying to map the expression levels of genes encoding enzymes. This tool colored EC accessions with green, red, and orange to represent up-regulated (UR), down-regulated (DR) and not significant (n.s.) gene expression, respectively. Considering Inositol Phosphate metabolism, 41 enzymes were identified. From those, 17 green nodes represent enzymes identified in soybean. Anchoring these sequences against the soybean genome, 165 probable non-redundant Ins loci were identified. The significant number of regions related to Ins genes reported in soybean is justified by its highly redundant (polyploid) genome. As expected, most of the unitags did not show significant variation, but the analyzed pathway included differentially expressed unitags. The tolerant accession showed a greater number of UR and DR candidates when compared with the sensitive accession, implying that such accession has a greater plasticity of transcripts responsive to the issued stress, which would allow a better matching (fine tuning) to the new condition. Our data point to the importance of the Ins availability for soybean root cell under dehydration (where many cellular processes require Ins), especially for plants which use free Ins to synthesize essential compounds, including those involved in hormonal regulation and stress tolerance.

Financial support: CNPq, CAPES.

Gene Regulatory Network Modeling for Mycelium-to-Yeast Transition of *Paracoccidioides brasiliensis*

Luciane Sussuchi da Silva¹, Célia Maria de Almeida Soares², Alexandre Melo Bailão², Clayton Luis Borges², Juliana Alves Parente Rocha², Juliano Paccez², Roosevelt Alves da Silva¹, Gustavo Goldman³, Luis Anibal Diambra⁴, Maristela Pereira²

Núcleo Colaborativo de Biossistemas, Federal University of Goiás¹; Departamento de Bioquímica e Biologia Molecular, Federal University of Goiás²; Departamento de Ciências Farmacêuticas, São Paulo University³; Centro Regional de Estudos Genómicos, National University of La Plata⁴

Recent advances in systems biology are focused on uncovering the mechanisms underlying the establishment and maintenance of cell states. Gene regulatory networks (GRNs) models have allowed to clarify complex interactions between large numbers of genes and identify genes involved in modulation of different cell phenotypes. Networks can be mapped as a graph whose nodes are associated with genes (or groups of genes) and edges depict the relation between the nodes. In this work, we have applied a GRN approach to analyze transcriptional data of *Paracoccidioides brasiliensis* cells going through mycelium-to-yeast transition. *Paracoccidioides* spp. (Pb) is a human pathogenic fungus responsible for paracoccidiomycosis (PCM), the most prevalent systemic mycosis in Latin America. The pathogenicity of Pb have been seen closed related to its dimorphic transition once the shift from mycelium to the yeast form is essential for infection. The experimental data (GEO, accession no.: GSE3238) monitor the gene expression of 4692 genes at several time points (5, 10, 24, 48, 72 to 120h) after a temperature shift (26° to 37°C) comprising the mycelium-to-yeast conversion. After preprocessing of control genes and biological replicates, genes with similar expression profiles were grouped in an optimum number of clusters (Nc=431). Intra-cluster averages of the expression level we used to model the GRN dynamics by a first order Markov model, where the future state depends linearly on the present state and on external perturbation. This approach allows the identification of regulatory genes mainly involved into the known phenotypic states and the environmental-cue effectors considering the transition between states. After analyzing the data, we described 20 genes with important regulatory activity for mycelium-to-yeast transition, whose three are up- or down-regulated by temperature. Among the genes coding for proteins with known function, we found proteins related to pro synthesis of cell wall and basal metabolism of the fungus.

Supported by: CNPq, CAPES, FINEP e FAPEG.

Integrating omics data from xylose-fermenting yeast using network dynamic modeling for bioethanol production

Lucas Miguel de Carvalho¹, Gabriela Vaz de Meirelles¹, Renan Pirolla^{1,2}, Leandro Vieira dos Santos^{1,2}, Gonçalo Amarante Guimarães Pereira¹ and Marcelo Falsarella Carazzolle¹.

(1) Laboratório de Genômica e Expressão, Instituto de Biologia-Unicamp, (2) BioCelere Agroindustrial LTDA-Campinas-SP, (3) Laboratório Dalton de Espectrometria de Massas, Instituto de Química-Unicamp.

Brazil is one of the biggest producers of ethanol in the world, a pioneer in the ethanol industry. However, the country is already facing a major limitation imposed by the first-generation ethanol production technology, which the sugarcane juice is converted by ethanol using industrial yeast *Saccharomyces cerevisiae*. Therefore, a new alternative has been proposed, called second generation, which is based on lignocellulosic residues of sugarcane (bagasse and straw) for ethanol production using recent methodologies for biomass desconstruction that generates soluble sugars, majority represented by glucose and xylose. One of the biggest challenges of this technology is the development of genetically modified industrial yeast that can not only produce ethanol from glucose as usual, but also from pentoses that represents 15% to 45% of the lignocellulosic material. Several works have developed xylose-fermenting yeast using different exogenous genes and genetic engineering approaches, but always resulting in very low yield and productivity mainly caused by unbalanced redox potential and metabolic bottleneck. The combination of omics data (transcriptomic, proteomic and metabolomic) and bioinformatics analysis is an essential step for a better understanding of this system. The objective of this work is to develop bioinformatics analysis for integration of omics data from xylose-fermenting yeast in different fermentation conditions. The omics data are being generated in our laboratory, complemented by public datasets and analyzed individually. After that, these data are submitted for protein-protein and metabolite-protein interaction networks using Integrated Interactome System, a web-based platform recently developed in our laboratory that gather novel identified interactions, protein and metabolite expression/concentration levels, subcellular localization and computed topological metrics, GO biological processes and KEGG pathways enrichment. Subsequently, the information is extracted from these networks and the dynamic modeling is applied to each node of the network. In first case of study, we used a public transcriptome dataset from industrial yeast *S. cerevisiae* during industrial-scale bioethanol production to perform protein-protein interaction network and simulations using expression profile for each gene. Basically, the bioinformatics pipeline developed until this moment converts metabolites networks from KEGG database (KGML format) to models in Petri Net (SBML format) and estimates the transition probabilities using transcriptome data, which can be dynamically modeled by the user using the software Snoopy. The software can produce Stochastic Petri Net models from regulatory and signaling networks. The results of the simulations were validated using experimental data as deletion and high gene expression in glycolytic network that alter the production of ethanol.

Design and Engineering of Synthetic Biological Systems for Medical Diagnosis

Francisco SANTOS SCHNEIDER¹, Alexis COURBET¹, Christophe NGUYEN¹;
Marina CARDIA JARDIM¹, Liyan HE¹; Laurence MOLINA¹; Liza FELICORI²;
Patrick AMAR^{1,3}, Franck MOLINA¹

¹Sys2Diag FRE 3690 CNRS/ALCEDIAG; ²Departamento de Bioquímica e Imunologia
Universidade Federal de Minas Gerais; ³LRI UMR 8623 Université Paris Sud CNRS

Bioinformatics and microfluidics provide useful tools for answering complex biological questions. On the one hand, bioinformatics allows us to design, model and simulate biological processes, resulting in the reduction of experimental tests. On the other hand, microfluidics provides miniaturization, automatization and portability, diminishing the required amount of samples and reagents. Together, they provide ways to create non-expensive devices which could be used in personalized medicine and provide new ways to probe, monitor and interface human physiopathology. In this work, we have applied engineering principles to design and build an artificial biological system, clinically compliant, that we have applied to Type 2 Diabetes Mellitus (T2DM) early diagnosis. Using two original softwares, we have constructed a synthetic biochemical network and created algorithms decision rules based on medical needs. To allow automatization and miniaturization of the process we have also conceived and produced microfluidic devices capable of producing highly stable and homogeneous double emulsion vesicles containing our biochemical networks. The diagnosis test-containing liposomes were used in biological samples (i.e. urine) and were capable of detecting T2DM early biomarkers, which could help medical decision. Thus, our methods based on a synthetic biology pipeline take advantage of both bioinformatics and microfluidic approaches, which enable simplified and efficient design of next generation devices for testing of human complex diseases.

An intuitive network-based approach to investigate clinical features among breast cancer subtypes

Fonseca, AL , De Souza, SJ

Universidade Federal do Rio Grande do Norte ,

Pós-graduação em Bioinformática , Núcleo Interdisciplinar de Bioinformática (BioME)

Breast cancer is a heterogeneous disease that covers a broad spectrum of pathologies with each subtype having particular characteristics, including: morphological, behavioral and molecular features. Furthermore, breast tumors have a variety of clinical outcomes, hormonal response and different therapeutic options. Nowadays, the 'omics' technologies have allowed the understanding of that heterogeneity and also their impact in patient prognosis. However, fundamental issues at clinical level remains unsolved, mostly regarding the association between molecular subtypes and treatment, such as the case of triple-negative breast cancer (TNBC). Furthermore, others issues associated with tumor aggressiveness and patient survival rate are not extensively investigated. To approach such problems, we create a network-based strategy for prioritization of genes and topological modules, called Triads. First of all, breast cancer data were collected from the TCGA consortium, which span four subtypes, Luminal A (440 samples), Luminal B (123), Her2 Enriched (37) and TNBC (113). In brief, our approach is based in two major principles, i) genomics data integration, and; ii) reduction of network complexity. For the data integration, TCGA data were summarized using the S-score method, which calculates for each gene a specific score, reporting alteration rate and their type (loss or gain-function). Finally, we apply a statistical correction to reduce the topological properties (nodes or edges) in each subtype network. As result, the network complexity reduction reached 50% and 30%, among the nodes and edges, respectively. Moreover, the nodes selected by our method presented a high proportion of extremely altered genes in each subtype. Furthermore, our findings report differential topological properties among the subtypes networks, for instance, exclusive nodes and edges in each network. To evaluate the biological meaning of these datasets, we carried out an enrichment analysis with the exclusive nodes against the KEGG and Gene Ontology databases. Finally, we analyze the iterated edges and their association with the survival outcome, through a Kaplan-Meier method. In addition, the frequency of survival edges in our networks seems to be significant when compared against full human PPI network. In conclusion, the development of network-based approaches is a powerful tool to understand complex diseases, mostly due to the ability of these methods to integrate biological data. We presented here, a flexible and intuitive approach able to reduces the network complexity and highlight relevant clinical features of different breast cancer subtypes.

SigNetSim : A web tool for modeling and analyzing quantitative biochemical networks

Vincent Noël , Marcelo S. Reis , Matheus H.S. Dias , Lulu Wu , Amanda S. Guimarães , Daniel F. Reverbel , Junior Barrera , and Hugo A. Armelin

LETA-CeTICS, Instituto Butantan, Brazil

Instituto de Matemática e Estatística, Universidade de São Paulo, Brazil

Instituto de Química, Universidade de São Paulo, Brazil

Molecular biology is experiencing a revolution, thanks to data acquisition becoming increasingly cheaper, and to the development of Systems Biology, an emergent research field that shows new ways to study such high-throughput data. However, biologists need a new generation of tools capable of handling large quantitative datasets, and also performing rigorous mathematical analysis of the kinetic models that may be derived from them. SigNetSim is a web tool coded in Python 2.7 using the Django framework, and Bootstrap as a graphical front-end. It is designed to perform most computer-intensive work server-side, which make it usable from most devices. It enables its user to easily create or edit quantitative models, using the latest standard to describe biological models (SBML L3V1). It is also compatible with the Hierarchical Model Composition package, which enable to describe models as a collection of modules, which are by themselves SBML models. SigNetSim can simulate models, both time series and steady states, and plot the results using interactive JavaScript libraries. It also includes some basic database to store experimental data, which associate a set of treatments to a set of observations. Database entries can later be used to simulate an experiment, using the list of treatments as initial conditions for the simulations. SigNetSim also enables the user to look for values of kinetic constants for the model which are able to reproduce the experimental data, using a parallel global optimisation method (simulated annealing). Finally, SigNetSim can perform dynamical analysis of the model, using exact methods thanks to a symbolic representation of the mathematical model, or other methods like numerical continuation techniques for bifurcation analysis. User can also use directly libSigNetSim, the core library of SigNetSim, from Jupyter notebooks in order to work directly on the models in Python, including the symbolic math version. Both the web interface and the library are available on GitHub, under free software license. It has already proved to be a very useful tool to work on quantitative models in our team. Thanks to its user-friendly web interface, even researchers and students that are inexperienced in programming can build, adjust and simulate models for both scientific and didactic purposes. We are currently adding support for model annotations, which allow to use standard formats to describe model components, and easily generate informative graphical diagrams.

This work was supported by grants #12/20186-9, #13/07467-1, and #13/24212-7 of the São Paulo Research Foundation (FAPESP) and fellowships from CNPq.

Hierarchical Model of the Ras-MAPK signalling pathway in mouse Y1 adrenocortical tumor cells

Vincent Noël¹, Marcelo S. Reis¹, Matheus H.S. Dias¹, Cecília S. Fonseca^{1,2}, Layra L. Albuquerque¹, Fábio Nakano^{1,3}, Junior Barrera^{1,4}, Hugo A. Armelin^{1,2}

¹LECC-CeTICS, Instituto Butantan, Brasil; ²Instituto de Química, Universidade de São Paulo, Brasil; ³Escola de Artes, Ciências e Humanidades, Universidade de São Paulo, Brasil; ⁴Instituto de Matemática e Estatística, Universidade de São Paulo, Brasil

The K-Ras-driven mouse adrenocortical tumor cell line Y1 displays a surprising association of phenotypic traits, i.e., high basal levels of activated K-Ras in starved cells and induction of cell cycle arrest upon stimulation by FGF2. In addition, ectopic expression of the dominant negative mutant Ras-N17 reduced activated K-Ras basal levels and eliminated cell cycle arrest by FGF2. We are working to uncover the molecular basis of this unexpected phenomenon by modeling the kinetics of the Ras-MAPK signalling pathway in Y1. We started by modelling the K-Ras molecular switch through a list of reactions, rate constants and initial concentrations, using the standard SBML format. With this simple model, containing only SOS to activate K-Ras, we were not able to reproduce the high basal level of activated K-Ras. This led us to add another GEF to our model, responsible for basal K-Ras activation. With this modification, our model was able to reproduce experimental data. We are currently planning to validate experimentally this hypothesis by knocking out GRP4, a GEF which has been shown to be very strongly expressed in Y1. We then moved to a Hierarchical Model representation using the SBML package comp, which enables us to represent models as a set of modules. We first included a simple module to describe the SOS activation by FGF2, which enabled us to represent the FGF treatment on our system. We then included another module representing how activated K-Ras induces the activation of the MAPK pathway. We finally added a module for describing Ras-N17 action, which reproduces observed behavior of Ras-N17 expression for both starved and FGF2-stimulated cells. We were able to start building a model to describe the unusual behavior of Y1 cells. Our model reproduces the behavior of K-Ras and MAPK activation in starved cells and FGF2-stimulated cells. We now need to improve it to include more treatments activating the pathway, and the consequences of these activations on the cell cycle control mechanism, to be able to describe and investigate the cell cycle blockage upon FGF2 stimulation. Our model is developed using standard representation, enabling collaborators to quickly start testing or improving our model. More precisely, the use of the SBML comp package enables us to build more modular models, which makes it easier when working with larger and larger models.

This work was supported by grants #12/20186-9, #13/07467-1, and #13/24212-7 of the São Paulo Research Foundation (FAPESP) and fellowships from CNPq.

Analysis and Mining Onco-targets Breast through Ontology

Edgar Lacerda de Aguiar¹, Lissur Azevedo Orsine², Carlos Alberto Xavier Gonçalves²,
José Miguel Ortega², Marcos Augusto dos Santos¹

*1 Bioinformatics and Systems Laboratory, Federal University of Minas Gerais, Brazil , 2 Biodata Lab,
Federal University of Minas Gerais, Brazil*

Studies indicate that by the end of 2016 more than 20 million patients will develop some type of cancer. Among the various types of cancer, breast cancer is the most impact among women and high mortality rate. Breast cancer has high biological heterogeneity, which implies a high diversity of molecular forms which are associated with distinct subtypes and distinct drug targets. This high range of variations in the biological entities involved in disease pathology impacts directly on diagnosis and treatment. Because of these facts this work aims to mine the possible genes related to breast cancer, from different databases (DBs), Cancer Genome Atlas (CGA), COSMIC and to relate the genes to database Gene Ontology (GO) with its molecular functions, biological processes, cellular components, thus inferring the main ontologies associated with breast cancer. Initially there was an interpolation of genes between BDs CGA and COSMIC after curated genes were mined and crossed with the top mutated genes of breast cancer. The genes were cured with the BDs UniProt and NCBI. With the UniProt valors Id of cured Genes there was a search for Ontologies in the GO database. The initial results were few satisfactory due to high specificity and high granularity of ontological terms. Better treatment of the data and a new methodological approach to the Ontological terms was necessary. The ontological terms of GoSlim, which are terms and healed with a median specificity were used. Through the analysis it was observed that many biological processes encountered are certainly associated with cancer, such as cell motility, cell adhesion, cell death, response to stress, immune system process and biosynthetic process. The molecular functions found are mainly engaged in activity in the DNA (DNA binding, nucleic acid binding transcription factor activity, transcription factor binding) and protein activity (enzyme binding, kinase activity, phosphatase activity, atpase activity, enzyme regulator activity). The analyzes show a significant correlation between the biological processes and molecular functions encountered: for example, signal transduction (biological process), and signal transducer activity (molecular function). This work proved to be essential to deepening of breast Oncogenes, as well as a better understanding of biological processes, molecular functions and their ontologies. These results will come to confirm the selection of the most critical processes for future studies of Oncogenes and Onco-targets.

Identifying metabolic processes shared among genome-wide association studies for reproductive phenotypes in bovine

Pablo A. S. Fonseca¹, Luíza A. F. Diniz¹, Fernanda C. Santos¹, Izinara C. Rosse¹,

Maria Raquel S. Carvalho¹

¹Departamento de Biologia Geral, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais – Brazil (MG)

The expressive improvement obtained in the production traits for bovines was followed by an increase in the frequency of reproductive disorders. Male reproductive traits have high heritability and a higher potential to segregate in the population, once male, generally, have largest offspring than females. Some association studies successfully identified genomic regions/genes associated with male reproductive phenotypes. Therefore, each one of these studies identified different candidate regions/genes, which difficult the interpretation of the biological meaning of the results, when evaluated together. An alternative to make this interpretation easier is the analysis of metabolic pathways in which each candidate genomic region/gene is involved, grouping them in larger biological units. The aim of the present study was to identify the metabolic pathways shared among studies that identified genomic regions associated with testicular size (3 studies) and noncompensatory fertility (2 studies) in bovines. In order to reach this aim, the genes mapped in those regions and the metabolic pathways, which each genes are related, were identified using the R packages biomaRt and reactome.db, respectively. From this analysis, 30 metabolic processes were found as shared among the three studies which identified genomic regions associated with testicular size. Moreover, 20 metabolic processes were shared between the two studies which identified genomic regions associated with noncompensatory fertility. The metabolic processes shared among the regions associated with testicular size in the three studies are related with cellular cycle control, regulation of translation and transcription, signal transduction, neuronal system and metabolism of lipids and proteins. These metabolic processes point to important components involved in the regulation of cell proliferation and metabolism of protein and lipids, contributing to the control of testicular morphology. Furthermore, supporting that the same processes, which regulate the increasing of milk production, could also regulate the development of reproductive traits. The 20 metabolic processes shared between the two studies, which evaluated the noncompensatory fertility phenotype, are related with the control of ionic channels, cellular interaction and activation of GABA receptors, metabolic processes directly related with the fecundation. The identification of these shared metabolic processes can help in the development of functional studies, leading to a better understanding of the biological basis of reproductive phenotypes in bovines. Moreover, once identified the most important pathways and biological markers, the development of breeding strategies aiming at the reduction of reproductive disorders frequency in the herds will be possible, consequently, reducing the economic losses caused by these phenotypes.

Gene network analysis of melanoma cancer development

Diego Vinícius de Castro Pereira, Andre Luiz Sena Guimarães, Fabiano Sviatopolk-Mirsky Pais

Faculdade Promove, Ave. João Pinheiro 164, Centro, Belo Horizonte, MG

Cancer is defined as a disorder characterized by uncontrolled cell growth and it is considered one of the leading causes of death worldwide. Cutaneous melanoma cancer, which affects melanocytes, has the highest growth rates in cancer in the last decade. Its appearance is somehow related to changes in specific genes and their own networks in the organism. Through a data mining approach, this study sought to identify genes that may potentially be involved with melanoma skin cancer development. Gene candidates were collected after extensive search at Pubmed database. As a control, genes related to the maintenance of a healthy skin where collected at GeneCards database. Then, the interactions between genes in each group were mapped, and a score related to this interaction was defined by the STRING database. A topological analysis of the interaction networks was performed with the Cytoscape software, which showed that both networks, melanoma and healthy skin, exhibit a power law behavior. Based on the score generated by STRING, the genes were clustered using K-means. Genes with the highest scores were identified as leaders. For the melanoma group, suggested genes leaders were: TP53, AKT1, JUN and STAT3 MYC. For healthy skin group, suggested genes leaders were UBC, TP53, JUN, AKT1, CREBBP, EP300 and SRC. At present, we are investigating the results in order to identify genes that could be closely related to gene leaders, according to the current methodology, without being previously described as related to melanoma development.

Financial support: FAPEMIG.

Ancestrality and evolution of genes related with apoptosis

Rayson Carvalho Barbosa, Carlos Alberto Xavier Gonçalves, José Miguel Ortega

Laboratório de Biodados, Instituto de Ciências Biológicas, UFMG

Apoptosis is a process of programmed cell death that occurs in multicellular organisms. It is an essential process to maintain the development of living beings acting to eliminate unnecessary or defective cells. During apoptosis, the cell undergoes changes in morphologic characteristics led by biochemical events. Such changes include cell contraction, loss of adhesion to the extracellular matrix and neighboring cells, chromatin condensation, internucleosomal cleavage of DNA, formation of apoptotic bodies, nuclear fragmentation, chromosomal DNA fragmentation, and total mRNA decay. The inhibitors of apoptosis proteins or IAP (Inhibitor of Apoptosis Protein) are molecules that exert their anti-apoptotic role through the ability to inhibit the activity of the effector caspase 3, 7 and 9 modulating the transcription factor NF -kb. We determined the Lowest Common Ancestor (LCA) for the genes on this system to investigate their origin along the evolution. We found that members of the Bcl-2 family (called anti-apoptotic regulators), responsible for inhibit apoptosis and prevent the release of cytochrome C, are present and conserved since Metazoa. This family also includes pro-apoptotic proteins Bax, Bid and Bak. Another protein important in cell cycle is DIABLO (generally referred as Smac/Diablo), that promotes caspase-9 for binding to IAPs and by removing their inhibition activities. We observed that these proteins have origin dated in Bilateria. Thus, we can conclude that the study of the evolution of proteins involved in the apoptotic process is important for your better understanding and can be used in combination with treatment and prevention of diseases related to disorders in this process.

GEN3VA: Aggregation and Analysis of Gene Expression Signatures from Related Studies

Gregory W. Gundersen, Kathleen M. Jagodnik, Holly Woodland, Nicholas F. Fernandez, Kevin Sani, Anders B. Dohlman, Peter Man-Un Ung, Caroline D. Monteiro, Avner Schlessinger, Avi Ma'ayan

Department of Pharmacology and Systems Therapeutics, BD2K-LINCS Data Coordination and Integration Center (DCIC) - Icahn School of Medicine at Mount Sinai, Fluid Physics and Transport Processes Branch - NASA Glenn Research Center, Center for Space Medicine - Baylor College of Medicine and Daylesford, the Fairway, Weybridge, Surrey, KT13 0RZ, United Kingdom

Genome-wide gene expression profiling of mammalian cells is becoming a staple of many published biomedical and biological research studies. Such data is deposited into data repositories such as the Gene Expression Omnibus (GEO) for potential reuse. However, these databases currently do not provide simple strategies to systematically analyze collections of related studies. Here we present GENE Expression and Enrichment Vector Analyzer (GEN3VA), a web-based system that enables the integrative analysis of aggregated collections of tagged gene expression signatures identified and extracted from GEO. Each tagged collection of signatures is presented in a report that consists of heatmaps of the differentially expressed genes; principal component analysis of all signatures; enrichment analysis with several gene set libraries across all signatures, which we term *enrichment vector analysis*; and global mapping of small molecules that are predicted to reverse or mimic each signature in the aggregate. We demonstrate how GEN3VA can be used to identify common molecular mechanisms of aging by analyzing tagged signatures from 244 studies that compared young vs. old tissues in mammalian systems. In a second case study, we collected 86 signatures from treatment of human cells with dexamethasone, a glucocorticoid receptor (GR) agonist. Our analysis confirms consensus GR target genes and predicts potential drug mimickers. GEN3VA can be used to identify, aggregate, and analyze themed collections of gene expression signatures from diverse but related studies. Such integrative analyses can be used to address concerns about data reproducibility, confirm results across labs, and discover new collective knowledge by data reuse. GEN3VA is an open-source web-based system that is freely available at: <http://amp.pharm.mssm.edu/gen3va>.

Metabolic pathways involved in bovine temperament

Fernanda C. Santos¹, Pablo A. S. Fonseca¹, Izinara C. Rosse¹, Luíza A. F. Diniz¹,
Maria Raquel S. Carvalho¹

¹*Departamento de Biologia Geral, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais – Brazil (MG)*

Bovine behavior is defined as the actions and reactions of the animal when in contact with humans and it is composed of many phenotypes. Among these, the most studied is temperament, which involves alertness, reactivity, physical movement, aggression, emotions, and even curiosity. This trait is very important because it reflects the fear felt by the animal in human presence, directly impacting on the animal-creator relationship and on the animal welfare. When compared to calm animals, the ones with bad temperament exhibit lower weight gain, lower reproductive efficiency and lower milk production. Besides that, these animals produce meat of lower quality, are more susceptible to diseases and are usually involved in work accidents. This way, animals with a bad temperament usually increase the costs of herd maintenance. As a complex trait, temperament is difficult to measure and, to date, only few studies on the genetics of temperament returned the same candidate regions, genes or QTLs. Many regions in different chromosomes have been identified as associated with temperament. A recent review of the literature on the genetics of temperament, reported 22 QTLs associated with temperament. These QTLs were obtained from 4 articles. All those results, difficult the interpretation of the biological causes for bad temperament, when analyzed together. In this context, the present study aims to identify the main biological pathways recovered by the four studies. A pathway was considered as shared, when it presented at least one gene from the same pathway in all the 4 articles. Additionally, an enrichment analysis was performed aiming at identifying the most enriched pathways recovered by the 4 studies. In order to reach these aims, the genes mapped in the candidate regions associated with temperament in these studies and the metabolic pathways, to which each gene is related, were identified using the R packages biomaRt and reactome.db, respectively. The enrichment analysis was performed using an Exact Fisher test in R. As a result, 14 pathways were shared among the 4 studies. Among these, we highlight Signal transduction, immune system, cell cycle and axon guidance. After Bonferroni correction, one shared pathway was still significantly enriched: the signal transduction. This pathway contains important genes that regulate the transmission of a molecular signal inside the cell, inducing cell modifications, recruitment of protein complexes, culminating on changes of expression patterns. Signal transduction pathway is essential to the functioning of synapses, being though a strong candidate pathway for behavioral traits.

