BIOINFO.IMD.UFRN.BR/NBF

ABSTRACT BOOK

NATAL BIOINFORMATICS FORUM

APRIL 15-17, 2019
GOLDEN TULIP PONTA NEGRA
NATAL/RN - BRAZIL

Sponsored by:

BiOME
Bioinformatics Multidisciplinary Environment
Centro Multiusuário de Bioinformática

Supported by:

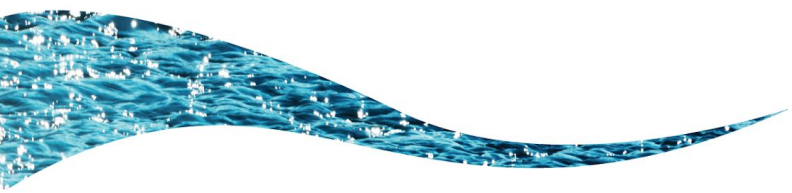
IMD Instituto Metrópole Digital
PPG BIOINFO UFRN
UFRN Universidade Federal do Rio Grande do Norte
biologiasistêmicadocâncer
2Bio Instituto de Bioinformática e Biotecnologia
CAPES
NATAL Convention Bureau
PPG BIOQ Programa de Pós Graduação em Bioquímica
Agilent Trusted Answers
NVIDIA

NATAL BIOINFORMATICS FORUM

APRIL 15-17, 2019
GOLDEN TULIP PONTA NEGRA
NATAL - RN - BRAZIL

ABSTRACT BOOK

Contents



Message from the organizing committee

Welcome to the ***Natal Bioinformatics Forum***! It's a great pleasure to have you all in Natal - Brazil, a city known by its paradisiac beaches and abundant sun, but also a city with great opportunities for those involved in technological and scientific fields, like Bioinformatics.

The NBF was designed to provide discussions on the perspectives and advances in bioinformatics research. The event will cover scientific contents like Genomics, Proteomics, Big Data, System Biology, and Evolution, but also will discuss the roles of Bioinformatics in the industries. These contents will be given by leading international and national scientists on the field in the form of keynote speeches, panels, and forums. During these three days of the event, we hope to construct an open environment allowing the interaction between students, entrepreneurs, and policymakers.

The organizing committee of NBF is comprised of professors and members of *Bioinformatics Multidisciplinary Environment* (BioME) which is part of *Digital Metropole Institute* (IMD) of *Federal University of Rio Grande do Norte* (UFRN). BioME hosts the Postgraduate Program in Bioinformatics (PPG-BioInfo) which provides academic training for master and doctorate degrees. Take your time to know more about BioME at https://bioinfo.imd.ufrn.br/.

In this first edition, NBF received a total of 118 participants, 26 international and national speakers, and 49 scientific works. The proposal for the next NBF is to occur in 2021. So, we hope to see you soon again in Natal!

NBF Organizing committee

Committees

Organizing Committee

SANDRO DE SOUZA
BioME - ICe/UFRN

JOÃO PAULO MATOS
BioME - DBQ/UFRN

RODRIGO DALMOLIN
BioME - DBQ/UFRN

GUSTAVO DE SOUZA
BioME - DBQ/UFRN

BEATRIZ STRANSKY
BioME - CTEC/UFRN

TETSU SAKAMOTO
BioME - IMD/UFRN

JORGE ESTEFANO DE SOUZA
BioME - IMD/UFRN

CÉSAR RENNÓ COSTA
BioME - IMD/UFRN

RENAN CIPRIANO
BioME - IMD/UFRN

GISELE TOMAZELLA
i2Bio

Support team and volunteers

MARIA SANTANA BRAGA
i2Bio

IARA SOUZA
BioME - PPg-Bioinfo/UFRN

CLÓVIS REIS
BioME - PPg-Bioinfo/UFRN

ANDRÉ FONSECA
BioME - PPg-Bioinfo/UFRN

GUILHERME ARAUJO
BioME - PPg-Bioinfo/UFRN

LEONARDO RS CAMPOS
BioME - PPg-Bioinfo/UFRN

DIEGO MARQUES-COELHO
BioME - PPg-Bioinfo/UFRN

TAYNÁ FIÚZA
BioME - PPg-Bioinfo/UFRN

LEVIR CHIANCA
BioME - BTI/UFRN

DANIEL GOMES
BioME - BTI/UFRN

RITA LOPES
BioME - BTI/UFRN

EULLE ARAÚJO
BioME - BTI/UFRN

Speakers

JESPEN OLSEN
University of Copenhagem, Denmark

MANYUAN LONG
University of Chicago, USA

MARTIN MORGAN
Bioconductor / Roswell Park Comprehensive Cancer Center, USA

PAUL VERSCHURE
University of Barcelona, Spain

PHILIP MAINI
Oxford University, UK

VINICIUS MARACAJÁ-COUTINHO
Universidad de Chile, Chile

ANDRÉ MAURICIO R SANTOS
Federal University of Pará, Brazil

AUGUSTO SCHRANK
Federal University of Rio Grande do Sul, Brazil

CÉSAR RENNÓ COSTA
Federal University of Rio Grande do Norte, Brazil

DANIEL DINIZ
Federal University of Rio Grande do Norte, Brazil

DAVID SANTOS MARCO ANTONIO
Grupo Fleury, Brazil

DIRCE CARRARO
A.C. Camargo Cancer Center, Brazil

EDUARDO EMRICH
Biominas, Brazil

GUILHERME CORRÊA DE OLIVEIRA
Instituto Tecnológico Vale, Brazil

GUSTAVO ANTONIO DE SOUZA
Federal University of Rio Grande do Norte, Brazil

JOSÉ IVONILDO DO RÊGO
Federal University of Rio Grande do Norte, Brazil

JOÃO PAULO MATOS SANTOS LIMA
Federal University of Rio Grande do Norte, Brazil

MAURO CASTRO
Federal University of Paraná, Brazil

PEDRO MÁRIO CRUZ E SILVA
NVIDIA, Brazil

RAPHAEL BESSA PARMIGIANI
IdenGene Medicina Diagnóstica

RENAN CIPRIANO MOIOLI
Federal University of Rio Grande do Norte, Brazil

RODRIGO JULIANI SIQUEIRA DALMOLIN
Federal University of Rio Grande do Norte, Brazil

RODRIGO ROMÃO DO NASCIMENTO
Federal University of Rio Grande do Norte, Brazil

SAMUEL XAVIER DE SOUZA
Federal University of Rio Grande do Norte, Brazil

SANDRO JOSÉ DE SOUZA
Federal University of Rio Grande do Norte, Brazil

SELMA JERÔNIMO
Federal University of Rio Grande do Norte, Brazil

Program

April 15th

9:50	**OPENING**
10:00	**CONFERENCE: *R / Bioconductor for the open-source analysis and comprehension of high-throughput genomic data*** Speaker: Martin Morgan, Roswell Park Comprehensive Cancer Center Chair: Rodrigo Dalmolin, UFRN.
11:00	**SYMPOSIUM: *Bioinformatics applied to Genomics/Proteomics*** ***Paving the way for the combined analysis of multiple proteomic datasets*** Gustavo A. de Souza, UFRN ***Bioinformatics/Genomics and the peopling of South America*** André M. Ribeiro dos Santos, UFPA ***The use of genomics tools to unravel the Amazonian biodiversity*** Guilherme Oliveira, Instituto Tecnológico Vale Chair: Tetsu Sakamoto, UFRN
12:30	**Coffee Break**
14:00	**CONFERENCE: *Modelling collective cell behaviour in biology*** Speaker: Philip Maini, Oxford University. Chair: Beatriz Stransky, UFRN.
15:00	**FORUM: *Bioinformatics in the Industry*.** Eduardo Emrich, BioMinas David Santos Marco Antonio, Grupo Fleury Raphael Bessa Parmigiani, IdenGene Medicina Diagnóstica Chair: Sandro J. de Souza, UFRN
16:15	**End of the day**

April 16th

9:20	**TECHNICAL LECTURE:** ***Integrated Biology analysis as a bioinformatics tool used to integrate multiomics data to early detect molecular alteration in cancer diseases*** Speaker: Maurício Marques, LCMS Product Specialist, Agilent Technologies Brasil
10:00	**CONFERENCE:** ***Dissecting oncogenic EGF receptor signaling in-vivo by quantitative interaction proteomics and phosphoproteomics*** Speaker: Jesper Olsen, University of Copenhagen, Denmark Chair: Gustavo A. de Souza, UFRN
11:00	**SYMPOSIUM:** ***Bioinformatics applied to Systems Biology*** ***Transcriptional Analysis based on PPI Networks*** Rodrigo Dalmolin, UFRN ***Chromatin accessibility and regulon activity patterns in breast cancer*** Mauro Castro, UFPR ***The challenges for modeling and simulation in bioinformatics*** César Renno Costa, UFRN Chair: Beatriz Stransky, UFRN
12:30	**Coffee Break**
13:00	**Poster Session**
14:00	**CONFERENCE:** ***De Novo Origination as Evolutionary Mechanism of Protein Diversity: The Standard and Evidence*** Speaker: Manyuan Long, University of Chicago. Chair: Sandro J. de Souza, UFRN
15:00	**FORUM:** ***Bioinformatics and Big Data*** Samuel X. de Souza, UFRN Vinicius Maracajá Coutinho, University of Chile Pedro Mário Cruz e Silva, NVIDIA Chair: Renan Cipriano Moioli, UFRN
16:15	**End of the day**
18:30	**Happy hour**

April 17th

10:00	**CONFERENCE:** ***The principles of neurorehabilitation: delivering brain repair at the confluence of brain theory, virtual reality and artificial intelligence*** Speaker: Paul Verschure, University of Barcelona, Spain Chair: César Rennó Costa
11:00	**SYMPOSIUM: Bioinformatics in Health** ***Circulating tumor DNA as a tool to monitor treatment response and antecipate tumor progression*** Dirce Carraro, AC Camargo Cancer Center ***The subvertion of the host metabolism in Leishmania infantum infection*** Selma Jerônimo, UFRN ***Assessing the structural impacts of cancer-associated mutations using new approaches*** João Paulo Matos, UFRN Chair: Lucymara Agnez, UFRN
12:30	**Light Lunch**
13:00	**Poster session**
14:00	**CONFERENCE:** ***Computational RNomics in Archaea: non-coding RNAs identification, biogenesis, conservation and lifestyle associations*** Speaker: Vinicius Maracajá Coutinho, University of Chile, Chile Chair: Sandro J. de Souza, UFRN
15:00	**FORUM:** ***Innovation in Natal: a Bioinformatics Perspective*** Ivonildo do Rego, UFRN Daniel Diniz, UFRN Augusto Schrank, UFRGS Chair: Rodrigo Romão, UFRN
16:15	**Closing**

ABSTRACTS

Observation:

- The abstracts are ordered by the Poster ID. The first digit in the Poster ID corresponds to the day in which the poster was presented (first or second day of the poster session).

Disclaimer:

- The contents of the following pages were produced using author-supplied copy. No responsibility is assumed for any claims, instructions or methods: it is recommended that these are verified independently.

The evolutionary interplay of human apoptosis, senescence, autophagy and genome-stability gene pathways

16 Apr
1:00pm
P101

Alana Castro Panzenhagen[1], Álvaro Oliveira Franco[1], Maikel Varal[1], Rodrigo Juliani Siqueira Dalmolin[2] and José Cláudio Fonseca Moreira[1]

[1]Centro de Estudos em Estresse Oxidativo, Programa de Pós-Graduação em Ciências Biológicas: Bioquímica, Instituto de Ciências Básicas da Saúde, Universidade Federal do Rio Grande do Sul, Porto Alegre, Rio Grande do Sul - Brasil.; [2]Bioinformatics Multidisciplinary Environment, Federal University of Rio Grande do Norte, Natal, RN, Brazil. Department of Biochemistry, Federal University of Rio Grande do Norte, Natal, RN, Brazil.

In the process of aging, cells must either adapt to stress or engage in programmed death. Genome stability pathways, i.e. DNA repair mechanisms and DNA replication, constantly maintain the cell by correcting eventual mutations and DNA damage. However, when residual damage accumulates, the cell must enter one of the following paths: senescence, permanent cell cycle arrest; apoptosis, genetic programmed cell death; or autophagy, damaged organelles turnover. Apoptosis, senescence, autophagy and genome stability pathways are therefore linked to cell cycle and are sometimes difficult to differentiate. Once these pathways are biologically related, the genes among them also overlap. Several studies have attempted to discern what mechanisms trigger different cell fates, but with little success. We propose that looking into these pathways relationships under the light of evolution and through network construction may help elucidate how they interact and whether they have similar origins or have instead converged functions through the cooptation of proteins. Our aim was to assess the entanglement between these pathways throughout evolution, unraveling which is more ancient and conserved and by which genes their integrated property emerged. Each group was extracted from described pathways in the Kyoto Encyclopedia of Genes and Genomes (KEGG) database. Next, this data was uploaded into the String-DB database in order to find their orthologous groups. The analyses were conducted with the geneplast R package, using a computational method to reconstruct the evolutionary scenario of each process in the eukaryotes. We have inferred the appearance of gene groups from their orthology information, in which we defined the root of different genes based on its orthologs distribution by searching the most probably reliable evolutionary root. We have also compared the conservative state of each process through the calculation of a plasticity index, in which we computed the evolutionary plasticity of the orthologous groups associated to the proteins in each pathway (apoptosis, autophagy, senescence, and genome stability related genes).These data were also plotted against the protein-protein interaction (PPI) networks of those pathways. The data from the last computations was analyzed through Kolmogorov-Smirnov curves for the rooting inference. This analysis considered the appearance of genes in function of the root (from most ancient to most recent). Moreover, the data concerning the plasticity inference was evaluated through Kruskal-Wallis test and Dunn's post hoc. The rooting analysis demonstrated that apoptosis seems to be the most recent of those pathways, while genome stability and autophagy appeared earlier in evolution. We also learned from

the plasticity analyses that senescence and apoptosis are more plastic than autophagy and genome stability pathways (but there is no significant difference between senescence and apoptosis). Additionally, the genome stability pathway is more conserved (less plastic) than all other pathways. These results agree with the rooting approach, since, generally, the more recent the process is, the more plastic its genes. Autophagy is usually involved in protein degradation, organelle turnover, and non-selective breakdown of cytoplasmic components. Those features are particularly important to any cell survival even when they are not in a colony. The turnover of cell constituents is of utmost importance for self preservation, a key component of evolution drivers. Genome stability genes are clearly well conserved and appeared before the other processes. This is in line with the function of those genes, which is maintaining the genetic code intact and minimize damage, which is primordial for every living cell. Furthermore, senescence seems to be placed in the middle, not recent or old, and relatively plastic. Our results suggest that senescence is a process gradually developing and that has not been stabilized yet. Senescence is a state of irreversible cellular arrest that can be triggered by a number of factors, such as telomere shortening or oxidative stress. This characteristic may also interfere in the analysis, since it is a very heterogeneous process. Our study is the first to use an evolutionary approach using orthologous groups and plasticity to evaluate this set of cell-cycle-related processes. The results point to clear differences in time of evolution and conservation of their orthologous groups, yet the relationship between their proteins should be studied further. Funding agencies: Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) Fundação de Amparo à pesquisa do Estado do Rio Grande do Sul (FAPERGS)

The role of vitamin D and sunlight incidence in cancer

16 Apr
1:00pm
P102

Alice Barros Câmara and Igor Augusto Brandão
Universidade Federal do Rio Grande do Norte

Background: Vitamin D (VD) deficiency affects individuals of different ages in many countries. VD deficiency may be related to several diseases, including cancer. Objective: This study aimed to review the relationship between VD deficiency and cancer. Method: We describe the proteins involved in cancer pathogenesis and how those proteins can be influenced by VD deficiency. We also investigated a relationship between cancer death rate and solar radiation. Results: We found an increased bladder cancer, breast cancer, colon-rectum cancer, lung cancer, oesophagus cancer, oral cancer, ovary cancer, pancreas cancer, skin cancer and stomach cancer death rate in countries with low sunlight. It was also observed that amyloid precursor protein, ryanodine receptor, mammalian target of rapamycin complex 1, and receptor for advanced glycation end products are associated with a worse prognosis in cancer. While the Klotho protein and VD receptor are associated with a better prognosis in the disease. Nfr2 is associated with both worse and better prognosis in cancer. Conclusion: The literature suggests that VD deficiency might be involved in cancer progression. According to sunlight data, we can conclude that countries with low average sunlight have high cancers death rate. New studies involving transcriptional and genomic data in combination with VD measurement in longterm experiments is required to establish new relationships between VD and cancer.

16 Apr
1:00pm
P103

Predicting Epstein Barr small-RNA targets in human neocortex

Annie Elisabeth Beltrão de Andrade, Víctor Serrano-Solís and Savio Torres de Farias

Laboratório de Genética Evolutiva Paulo Leminsk, Departamento de Biologia Molecular, Centro de Ciências Exatas e da Natureza, Universidade Federal da Paraíba, João Pessoa, Brazil

A microRNA (miR) is a small RNA sequence, comprising 20 to 24 nucleotides, encoded from non-coding RNA genes (ncRNAs). They exert an important role in post-transcriptional gene regulation, matching homologous sequences at mRNAs and silencing the translational process. Besides the canonical biogenesis of miRs, others types of smallRNA can act as the classic miR, but they arise through metabolic processes of other classes of RNAs and eventually enter the miR pathway. The miR precursor when processed by the enzyme DICER, can give rise to isoforms, besides the mature miR, called isoMIRs, which may act as the mature form. Also, molecules derived from the maturation of tRNAs can accumulate in the RISC effector complex and disrupt the pathway. This occur because hairpin molecules can be recognized by enzymes of the canonical genesis and entering the metabolic pathway of miRs. The understanding of the biogenesis of these miRs and other sRNAs, has emerged as important factors both at regulation of physiological and pathological processes.

DNA-genome virus are utilized as experimental subjects for understanding such processes. The aforementioned ncRNA are present at viral infection of herpesviruses. The Epstein-Barr (EBV) is an ecumenical human pathogen of the Herpesviridae taxonomic family that regulates both viral replication and host response through miRs. In addition, the EBV is known to be associated with the development of autoimmune cancer and neuropathologies. The molecular mechanisms are still unknown, however, many studies have demonstrated that the expression of microRNAs correlates with the development of such pathological processes by regulating the host's immune system and viral expression for a lifespan latency. In this context, we sought to understand through computational tools how EBV makes use of miRs and other small ncRNAs to modulate mechanisms of hosts neurological physiology. For viral miR predictions we used the StructRNAfinder program, a tool for prediction and functional annotation of ncRNA through covariance models of secondary structure. After ncRNAs were characterized, we proceeded with the target gene selection of the neocortex area, using the public database Human Brain Transcriptome (HBT) that aggregates information of cerebral gene expression at tissue level. Then, we used IntaRNA program to predict miR-mRNA interactions. A main characteristic evaluated with this tool was the Free Minimum Energy value of the predicted interactions. Low free energy values are related to stability of molecular complexes, so these values were chosen because they better represents the biological systems. The result obtained from the HBT were 17 genes related mainly

to immune system, neurotransmission, signal transduction, transcriptional regulation, cell adhesion, cytoskeletal remodeling, hormone receptors and chromatin remodeling; some genes are as well related to neurodegenerative diseases and cancer development. We predicted 55 ncRNA for the EBV genome. Best detected interactions occurred over 8 of the 17 selected neocortex mRNAs. Interactions between mRNAs and EBV miRs showed low interaction energies, with values ranging from -30 to -62 kcal/mol. According to the literature, free energy lower than -25 kcal/mol is considered a stable energy value. These energy values were obtained from the interaction between the viral miRs and 8 genes from the human neocortex region. These genes participate in important pathways in signal transmission (CABP1, HTR3B and NUAK1), regulation of gene transcription (PKNOX2, POU6F2), cell adhesion (MKX) and chromatin remodeling (SATB1, SATB2). The neocortex presents intense neuronal activity, due to its complex role of information transmission related to perception, learning and memory. Disruptions in the proper functioning of the neocortex are related to neurodegenerative disorders, decline in cognitive abilities and dementia. The understanding of molecular interactions in the EBV infection on neuronal tissue is crucial for elucidate how and why these neuronal complication arises in determined percentage of cases.

The endocannabinoid system originated in the ocean, while the biosynthesis of phytocannabinoids in the period of dinosaurs

16 Apr
1:00pm
P104

Beatriz Moura Kfoury de Castro[1],Tetsu Sakamoto[2] and José Miguel Ortega[1]
[1]UFMG;[2]UFRN

The plant Cannabis sativa, known as marijuana, is a species of flowering plant in the family Cannabaceae. Cannabis produce phytocannabinoids in the secretory cavity of the glandular trichomes, which largely occur in female flowers and in most aerial parts of the plants, predominantly represented by a group of C21 or C22 (for the carboxylated forms) terpenophenolic compounds. Its predominant compounds contain Δ9-tetrahydrocannabinolic acid (THCA), cannabidiolic acid (CBDA) and cannabinolic acid (CBNA), followed by cannabigerolic acid (CBGA), cannabichromenic acid (CBCA) and cannabinodiolic acid (CBNDA). In human history, marijuana has been known since ancient China around 5000 years ago, where extracts of the plant were used for relief of cramps and pain. Wherefore, cannabinoids represent the most studied group of compounds, mainly due to their wide range of pharmaceutical effects in humans, including psychotropic activities. Currently, it is known that natural cannabinoids are also produced by the human body and other animals as endocannabinoids. These cannabinoids are endogenous lipids derived from polyunsaturated fatty acids. The two most known endocannabinoids, Anandamide (AEA) and 2-arachidonoylglycerol (2-AG), are both derivatives of arachidonic acid. These molecules found in all mammalian tissues. Together with cannabinoid receptors (CBs) and the enzymes responsible for the synthesis and degradation of the endocannabinoids comprised the endocannabinoid system (ECS). Its widespread neuromodulatory system plays important roles in central nervous system (CNS) development, synaptic plasticity. They bind to G protein coupled receptors (CB1 or CB2), as well as the phytocannabinoids, which mediate their action and being part of the ECS. Their role as retrograde messengers suppressing transmitter release in a transient or long-lasting manner, at both excitatory and inhibitory synapses, is now well-established. These lipid messengers can regulate several neural responses. Activated CBs suppresses the release of neurotransmitter in two ways: first, by inhibiting voltage-gated Ca2+ channels, which reduce presynaptic Ca2+ influx and second, by inhibiting adenylyl cyclase and the subsequent cAMP/PKA pathway. Cannabinoid receptors participate in the regulation of many physiological processes, such as emotion, reward, cognition, inflammation and reproduction. To retrieve genes associated with cannabinoids, both in its biosynthesis in Cannabis sativa, in its endogenous cannabinoids system in humans, scientific papers were collected describing the molecular and chemical biology of cannabinoids biosynthesis and the molecular biology of endogenous system their receptors. They were analyzed manually. After this selection, the metabolic pathways were constructed with manual curation. To analyze the evolutionary origin of genes in pathways, their sequences were retrieved from

the UniProtKB database and their orthologs from other species were retrieved using TaxOnTree software produced by our group. We then determined the lowest common ancestor (LCA) between species with representative proteins in the cluster of orthologs therefore inferring the origin of the gene. As a result, we were able to find 6 genes linked to the biosynthesis in plants and 152 genes linked to the endocannabinoid system in humans. Inference from gene origin showed that most of the genes that form the cannabinoid biosynthesis pathway appeared in the more recent clades of the evolutionary history of plants, the clade Mesangiospermeae (core angiosperms). In the human system pathway, inference has shown that most genes originated in the early Cambrian, in gnathostomata, including cannabinoid receptors. However, it seems that the endogenous cannabinoids and their degradation system are more ancient than the present receptors. Thus, they might already have some function in Bilateria. Presumably, other cannabinoid receptors could play the role in the system in these animals. It also seems that biosynthesis in plants was likely to have originated later in the presence of animals on Earth. Our data showed that the biosynthesis of phytocannabinoids probably originated at later of the Jurassic (150Ma). Period dominated by dinosaurs in which the first birds also appeared during the Jurassic, having evolved from a branch of theropod dinosaurs. Other major events include the evolution of therian mammals, including primitive placentals.

16 Apr
1:00pm
P105

Systems biology-based analysis of lead-treated human neural progenitor cells

Clóvis F Reis[1], Iara D. Souza[1], Diego Morais[1], Raffael Azevedo[1], Danilo Imparato[1], Rita MC de Almeida[2] and Rodrigo Juliani Siqueira Dalmolin[1]

[1]UFRN; [2]UFRGS

Lead is an essential raw material and is largely used in industry. But it also is very toxic to biological systems with important implications for human health. This heavy metal virtually affects all human systems, including nervous system impairment, peculiarly during development, leading to neural damage. Lead interaction with calcium and zinc-containing metalloproteins broadly affects cellular metabolism since these proteins are related to intracellular ion balance, activation of signaling transduction cascades, and gene expression regulation. In the present work, we applied the transcriptogram methodology in an RNA-seq dataset of human embryonic-derived neural progenitor cells (ES-NP Cells) treated with 30 μM lead acetate for 26 days. Transcriptogramer is a non-supervised system biology-based method designed to identify functionally associated gene groups differentially expressed in case-control designed experiments. Taking data from all 26 days of lead exposure, we defined two groups of similar samples based on a clustering algorithm. Those groups of samples were used as input for the transcriptogramer pipeline that could identify 11 clusters in both time-intervals with distinct expression patterns. Up-regulated clusters were small and presented low connectivity networks with few enriched terms, and down-regulated clusters presented big and tightly connected networks with a long list of enriched terms. Taking the enriched terms together, we observed early downregulation (from day 3 to 11) of several cellular systems involved with cell differentiation, such as cytoskeleton organization, RNA and protein biosynthesis, and interference in all layers of gene expression regulation, from chromatin remodeling to vesicle transport for long treatment times (from day 12 to 26). Considering ES-NP Cells are progenitor cells which can originate other neural cell types, our results suggest that lead-induced gene expression disturbance might impair cells ability to differentiate, therefore influencing ES-NP cells fate.

Diving in the frequency domain to seek for attractors of Boolean gene regulatory networks.

16 Apr
1:00pm
P106

Dhiego Souto Andrade and César Rennó-Costa
UFRN

Modelling gene regulatory networks as Boolean networks allow in silico examination of the metabolic kinetics that drive cellular behavior. However, the intrinsic nonlinearity of such Boolean networks hinders the prediction of future network configurations, and therefore the cell fate, besides the identification of the specific roles of individual components. A common approach is to look for stable states, known as attractors, that are poorly influenced by perturbations such as topology alterations and different initial states. The existence of attractors indicates a reduction in the dynamics entropy which ultimately support stable cellular computation. Therefore, attractors can be linked to specific phases of cellular metabolic process, such as the cell cycle and differentiation and can bring forth the association between specific nodes to cellular phenotype in genetic expression data. Thus, one important step in the study of Boolean networks is the identification of such attractor states. However, especially for large networks, the diversity of initial conditions and possible perturbations, makes it a difficult computational problem to identify the attractors. Here we describe a method to identify attractors in dynamics complex networks based on the analysis of network dynamics in the frequency domain. The premise is that components within the same attractors should oscillate with similar frequency components, they should become evident in a representation in the frequency domain. A hierarchical clustering applied on the frequency matrix, shows groups of nodes gathered by their frequency components, that, based on our assumption, are the groups of nodes that form the attractors. To test this assumption, we used a well-known model of cell cycle with 10 states. This model shows a synchrony of 7 nodes and the absence of activity of 3 nodes during cell cycle Boolean dynamics. We expected and obtained, as a result, a frequency domain matrix with one high activity cluster of size seven, that represents the unique synchrony of these attractor nodes, corroborating with our assumption. An investigation of the number and size of clusters fathered by larger networks were performed using Barabási-Albert networks, which carry features observed in biological networks. We simulated networks of sizes from 50 to 200 nodes. Our first results showed that the bigger the network, the bigger is the size difference between the clusters, that is, in average, networks with sizes near 50 resulted in clusters with almost the same size, and networks with sizes near to 200 resulted in a matrix with one big cluster followed by other small ones. This analysis strongly suggests that large complex networks tend to have a stronger force of synchronization of small attractors into a bigger one. For future work we will use gene expression data of patients with cancer and try to identify specific attractors nodes important to the attractors stability and attractors linked to subtypes of cancer.

16 Apr
1:00pm
P107

SOFTWARE FOR THE OPTIMIZATION OF GERIATRIC EVALUATION

Éberte Valter da Silva Freitas and Remerson Russel Martins
Universidade Federal Rural do Semi-Árido

Increasing population aging and multi-morbidity lead to increased demand for health care and increased spending. Emerging technologies, specifically electronic and mobile healthcare platforms, can be used to prevent illness and help seniors and healthcare professionals maintain a healthy and functional life. This study developed and evaluated a m-health platform for Comprehensive Geriatric Assessment (AGA) to be used in primary care, providing the professional with indications of preventive action and, consequently, delaying the establishment of damages to the health of the elderly. It is an exploratory research in the modality of technological development, with a qualitative-quantitative approach, divided in three stages: selection of AGA tests; design process; and, implementation of the conceptual model. In the first, a narrative and integrative review of the literature on the evaluation scales of elderly people employed in Brazil and in the world was done; In order to do this, the filtering and selection of the scales to compose the mobile platform, within the universe of tests obtained in the previous stage, by two specialists (1 in neurology and 1 in health of the elderly person); and finally the evaluation of professionals in the area of collective health to refer to the selected scales considering their feasibility and sensitivity for basic care. The definition of the conceptual model took place through the study of the functionalities of the product, its requirements and modeling it through UML diagrams. After the implementation of the conceptual model, a MARS form (1 = inadequate and 5 = excellent) was evaluated along with 11 professionals from the health area of primary care and 11 from the ICT for the formation of feedback of suggestions. The study was submitted to the CEP under protocol 01062818.2.0000.5294. The review resulted in 29 scales, of which 9 were selected for the m-health evaluation library. The modeling of the system defined the need for an initial anamnesis; contain physical, mental, and functional exams in your library; to raise problems and to identify priorities according to the results, and thus to suggest an option of intervention, being also, defined the prerequisites, the classes and the relationships of the subjects involved. The beta version of the platform was evaluated positively (general mean = 3.94 $\pm$ 0.95) and higher than 3 in all dimensions (engagement = 3.81 $\pm$ 0.90, functionality = 4.39 $\pm$ 0.65, aesthetics = 3.93 $\pm$ 0.85 and information = 3.74 $\pm$ 1.07), obtaining Krippendorff's Alpha equal to 0.2231. The evaluation by areas was also positive in both the biomedical area and the ICT, being better evaluated among the first ones (respectively, 3.96 $\pm$ 0.88 and 3.67 $\pm$ 1.03 in the engagement dimension, 4.45 $\pm$ 0.56 and 4.33 $\pm$ 0.69 in the functionality dimension, 4.15 $\pm$ 0.90 and 3.72 $\pm$ 0.91 in aesthetics, and 3.9 $\pm$ 0.98 and 3.57 $\pm$ 1.08 in dimension information). The m-health platform for Broad Geriatric Assessment to be used in

basic care developed has a high degree of quality according to ICT professionals and potential users.

QUANTUM BIOCHEMISTRY IN ZINC-DEPENDENT METALLOPROTEIN: A STUDY IN PORPHOBILINOGEN SYNTHASE

16 Apr
1:00pm
P108

Emmanuel Duarte Barbosa, Umberto Laino Fulco, José Xavier Lima-Neto and Viviane Souza do Amaral

UFRN

Metalloproteins are challenging objects when it comes of compromise between accuracy and com- putational feasibility. At present, quantum mechanics methods have been the key to describe the intrinsic relationship between metals ions and the relevant residues in proteins. Here, we report the calculated interaction energy curves for ion zinc and porphobilinogen synthase (PBGS) amino acid obtained by molecular fragmentation with caps conjugate (MFCC) and full system schemes. For accomplishment of quantum calculations were employed two exchange correlation functional and for expansion of Kohn-Sham electronic orbitals was used three basis-set models. In order to ob- serve the surrounding effect each step of MFCC, we applied four different dielectric constants. Our results point to significant differences in accuracy and computational cost among the calculation models. In addition, we describe the energetic mapping of the relevant residues that contribute to the maintenance of the zinc ion in the catalytic pocket of PBGS.

Comparative genomics of cellulose degradation mechanism across all the domains of life

16 Apr
1:00pm
P109

Fenícia Brito Santos[1], Tetsu Sakamoto[2] and J. Miguel Ortega[1]
[1]UFMG; [2]UFRN

Cellulose is the most abundant terrestrial biopolymer on Earth, along with hemicellulose and lignin. The latter is the second most abundant recalcitrant natural polymer, together they are the main components in the plant cell wall. The great world demand for energy and fossil fuels, generate serious environmental problems and might lead to the scarcity of nonrenewable resources. For this reason, clean energy sources have been an option nowadays and has drawn attention of the biofuel industry. Biofuels are a clean energy alternative derived from the decomposition of plant material from various sources, such as agricultural residues, forestry wastes, and energy crops. The conversion of lignocellulosic biomass to ethanol involves physical, chemical and physiochemical pretreatment to release the cellulose and hemicellulose. In sequence, enzymes from fungi and/or bacteria perform the breakdown of those polymers. Fungi are well known as lignocellulose decomposers, they have a great enzymatic arsenal for degrading those polymers. Although, other organisms have also been reported to have some enzymes associated to the lignocellulose degradation, how much of that arsenal those other species hold? Here, we searched the genomes of 6873 species on the Reference Proteomes database for the set of enzymes for lignocellulose degradation. Using the using the software TaxOnTree to performed phylogenetic and taxonomic distribution analysis, we were able to identify which organisms have which proteins and describe their potential mechanism for decompose the lignocellulose polymer. Our results show that in the Archaea domain the species from the Halobacteria class have few enzymes related to the hemicellulose degradation. Those are free living organisms that must benefit to the interaction with other lignocellulose degrading species. The degradation of cellulose on the bacteria domain are restrict to the enzymes Beta-glucosidase and Cellobiose dehydrogenase, suggesting that they can only break the end of the cellulose polymer. The majority of the bacteria species lack the full set of enzymes for the breakdown of hemicellulose, although some groups, e.g. the Actinobacteria and Bacteroidetes, have almost all enzymes to decompose this polymer. Most of the orders from the fungal phyla, Ascomycota and Basidiomycota, have a complete arsenal for the degradation of cellulose and hemicellulose, which explain why they are known as the best lignocellulose decomposers, and the most commonly used on the industry. Some metazoans, e.g. arthropods, mammals, annelids and mollusks, also have some enzymes related to the cellulose and hemicellulose breakdown. Similarly, to the Halobacteria case, those organisms must have acquired those genes for horizontal gene transfer, due to environmental interactions and their nutritional habits. Plants also have some of the enzymes to decompose cellulose and hemicellulose. Those enzymes, are important on

the process of synthesis and degradation of their cellular wall. Finally, we can conclude that the degradation of cellulose and hemicellulose are spread across all domains of the tree of life. Those enzymes are important for their nutritional habits. The most spread enzyme is Beta-glucosidase and Cellobiose dehydrogenase, even among those that do not hold the complete set of enzymes. This indicate that those organisms are able to break the polymer in some level, allowing them to use the cellulose as a carbon source. The hemicellulose is composed by different monomers, therefore the species that have a subset of the enzymes for it decomposition are able to use it as a carbon source. Our results show that bioinformatics analysis may be a powerful tool for the bioprospection of species for studies on the degradation of lignocellulose.

Integrated Multiomics Pipeline for prediction of Associated Risk Variants in Enhancers

16 Apr
1:00pm
P110

Francisco Eder de Moura Lopes[1], Suzana Pôrto Almeida[1], Rafael Mendonça Colares[1], Raul Victor Medeiros da Nóbrega[2], Shara Shami Araújo Alves[2], Saul Gaudencio Neto[1], Leonardo Tondello Martins[1], Ana Cristina de Oliveira Monteiro Moreira[1], Antonio Carlos da Silva Barros[2], Pedro Pedrosa Rebouças[2], Victor Hugo Costa de Albuquerque[3] and Kaio César Simiano Tavares[1]

[1]Laboratório de Processamento de Imagens e Simulação Computacional, Instituto Federal de Ciência e Tecnologia do Ceará, Fortaleza, Ceará, Brazil; [2]Programa de Pós-Graduação em Informática Aplicada, Universidade de Fortaleza, Ceará, Brazil; [3]Laboratório de Biologia Molecular e do Desenvolvimento, Universidade de Fortaleza, Fortaleza, Ceará, Brazil

INTRODUCTION AND OBJECTIVES In the biology systems era, a software that combines information from different omics databases and also does data mining have become increasingly useful into suggesting models for experimental validation. At the same time, a growing number of GWAS (Genome-wide Association Studies) are demonstrating that SNPs (Single Nucleotide Polymorphisms) occur inside non-coding regions most often than in genes. Therefore, the urgency of reverse genetics investigations in these regions is evident. However, if cis-regulatory elements (e.g. enhancers) could be thousands of nucleotides away from its respective regulated genes, the definition of a start point in the elucidation of these relationships are challenging. The answer may be exactly in multiomics. In an innovative work, Soldner et al., (2016) employed a new approach to find cis-regulatory elements associated with Parkinson's Disease from integrated multiomics data. Their results demonstrated the in vitro and in vivo role of a particular enhancer and its related transcription factors in a pathway that improve the amount of SNCA, a known physiopathology feature of Parkinson's Disease. Nevertheless, they used a non-automatized and laborious protocol to search for candidate enhancers. Inspired by this work, our aim was to build an integrated pipeline considering a similar approach that could find candidate enhancers in an easier and faster way.

MATERIAL AND METHODS We developed a new Python integrated pipeline. In the first step, we ask for a personalized SNPs set with reference ID number for all SNPs. From this, we request the essential details for each SNP directly from dbSNP database (https://www.ncbi.nlm.nih.gov/projects/SNP/). Secondly, the SNP's genome localization is used to investigate epigenomic marks in its region, searching in RoadMap Epigenomics Database (http://www.roadmapepigenomics.org/) for H3K4m1, H3k4m3, H3K27ac or the states models on a particular tissue or cell type. Hence, the user can choose only the SNPs inside enhancers for the next step. The last pipeline procedure is the transcription factors (TF) module. Here, we used one of the most validated tools designed for the search of TF binding sites: FIMO − Find Individual Motif Occurrences (http://meme-suite.org/doc/fimo.html). Automatically, our script makes

the analysis of differential TF, optimizing the most tedious and subject to human errors phase. In the end, we configure the tool to returns a bar chart illustrating the number of differential TFs that binds in each SNP region and a table of these detailed results. Finally, we have performed a validation round using the 463 Soldner's SNPs for Parkinson's Disease.

RESULTS AND DISCUSSION We developed a new Python script that starts from a personalized SNPs set, which may be associated a complex trait of interest, and results in candidate enhancers and transcription factors. In other words, we successfully did the integration of the genomics, epigenomics and transcriptomics data and applicated a regulomics prediction to it. In the validation round, we found some distinct results to the Soldner's manual approach, but we can conclude that it happened because of intrinsic differences to the method used by FIMO and it does not represent any pipeline integration error.

CONCLUSIONS Now, it looks like a powerful tool for its original purposes, to find candidate enhancers and to suggest its subjacent regulatory mechanisms. Currently, efforts to convert these pipeline in a user friendly application are in progress.

REFERENCE SOLDNER, F. et al., Parkinson-associated risk variant in distal enhancer of α-synuclein modulates target gene expression. Nature, v. 533, n. 7601, p. 95-99. 2016.

Funding Agency: CNPq - Conselho Nacional de Desenvolvimento Científico e Tecnológico CAPES – Coordenação de Apefeiçoamento de Pessoal de Nível Superior

COMPREHENSIVE BIOINFORMATICS ANALYSIS OF THE IMMUNE MECHANISM IN BREAST CANCER SUBTYPES

16 Apr
1:00pm
P111

Genilda Castro de Omena Neta, Paula Jordana da Costa Silva, Rafhaela Albuquerque Feitosa, Liliane Patrícia Gonçalves Tenório, Ricardo Jansen Santos Ferreira, Carolinne de Sales Marques, Aline Cavalcanti de Queiroz and Carlos Alberto de Carvalho Fraga

UNIVERSIDADE FEDERAL DE ALAGOAS

Introduction and Objectives: Breast cancer is the most common cancer in women worldwide. The biological characterization of breast cancer subtypes has clinical implications, impacting the treatment and, consequently, the survival after diagnosis. An important development in tumor immunology was the identification of highly diverse tumor-infiltrating leukocyte subsets that can play strikingly antagonistic functions. While earlier works had showed tumor cell-derived factors associated with inhibition of the local immune response, the past few years have demonstrated a dramatic contribution of leukocytes themselves to this "pro-tumor" environment. The present study aims to compare the cells of the inflammatory infiltrate in the different subtypes of breast cancer using different computational tools. Material and methods: 1257 patients were analyzed, the data extracted from the Cancer Genome Atlas (TCGA), a database of free access available online. We used TCGAbiolinks, R / Bioconductor software and the TCGAbiolinksGUI interface to download genomic and clinical data from normal and solid tumor tissues. We retrieved level data for raw count mRNA (Illumina HiSeq 2000). Co-expressed upregulated and downregulated Differentially Expressed Genes (DEGs) from the gene expression profiles were combined and identified with a Venn Diagram 2.1.0. The common DEGs were analyzed using Database for Annotation, Visualization and Integrated Discovery version 6.8, an online platform that provides a comprehensive set of functional annotation tools for investigators to understand biological meaning behind large list of genes. The value $P<0.05$ was considered to indicate statistically significant difference. Twenty-three leukocyte types and subtypes were compared in addition to the immunological response genes (Th1, Th2 and Th17) and a list of neuropeptides and their receptors. The data were confirmed by the analysis of the immune-oriented view and analysis of the Prediction of Clinical Outcomes from Genomic Profiles (iPRECOG). A survival analysis was performed through the PRECOG platform. Statistical analyses involving cell count number were performed using GraphPad Prism 7 software. A two-way ANOVA and a two-tailed unpaired t test were used to compare the means between groups. Results and Conclusions: Leukocyte cell analysis showed that macrophages are directly associated with different subtypes of breast cancer. M1 macrophages, Th1 and Th17 responses are increased in the Basal and Her2 subtypes, while both Luminal A and Luminal B showed M2 macrophage and Th2 response activation. Although the results showed M1 macrophage, Th1 and Th2

activation in Basal and Her2 breast cancer subtypes, we observed that M2 macrophage cell count number are increased in all breast cancer subtypes when compared with M1 macrophage. When we evaluated a list of neuropeptides and their ligands, we found thirty three neuropeptides mRNA expression increased in Basal and Luminal B subtypes. DAVID analyses confirmed the data indicating that M2 macrophages maybe are associated with neuropeptides expression, acting as potent cellular growth factors. Survival analysis showed neuropeptides are associated with poor survival in Basal and Her2 subtypes. In conclusion, we found Th1 and Th17 phenotype increased in Basal and Her2 breast cancer subtypes, while Th2 response is activated in Luminal A and Luminal B. M2 macrophage are increased in all cancer subtypes, and they are associated with neuropeptides mRNA expression, having survival impact in Basal and Her2 breast cancer subtypes.

Evolutionary root inference of orthologous groups related to eukaryotic essential genes

16 Apr
1:00pm
P112

Iara D. Souza, Clóvis F. Reis and Rodrigo J. S. Dalmolin
BioME/UFRN

Introduction and objectives: Genes crucial for cellular maintenance are called essential genes. They participate in cell core functions and deleterious mutations on those genes result in lethality. In multicellular organisms, deleterious alterations on essential genes produce a spectrum of phenotypes, from the impairment of fertilization process, disruption of fetal development, up to loss of reproductive capacity in adult individuals. We suppose that considering the importance of processes regulated by essential genes, they arose earlier during the evolutionary history and were conserved in many organisms. Here we aim to identify the essential genes in eukaryotic organisms and estimate their evolutionary root. Materials and methods: We retrieved phenotype information of mutant genotypes from specialized model organism databases. Based on whether the mutant genotype can be lethal, we classified genes in essential and non-essential. In addition, according to which developmental stage the lethality occurred, we have categorized mouse essential genes into three groups: early lethality, intermediate lethality and late lethality. We have identified the orthologous groups belonging to essential and non-essential genes with STRING database (v10.5). Using the Geneplast R package, we inferred the evolutionary root of essential and non-essential genes. Results and conclusion: Essential genes, in average, have a greater ancestrality index than non-essential in all organisms considered. Similarly, the root distribution of essential genes categories in mouse shows a prevalence of early lethality genes in primitive roots. Therefore, we concluded that essential genes emerged earlier than non-essential genes in evolutionary history.

16 Apr
1:00pm
P113

An interface for combined analysis of multiple proteomic datasets

Jade M.E. Gomes, Jose E. Kroll and Gustavo A. de Souza
Bioinformatics Multidisciplinary Environment, UFRN Brazil

In the past decade, proteomic analysis had evolved from a tool which could barely investigate hundreds to a couple thousand proteins in a given sample to a tool able to detect the complete transcriptome of a cell. However, due to many differences in mass spectrometry technology, sample preparation, quantitation approaches and bioinformatic tools used to investigate the data, the comparison of datasets which were generated independently is still a challenge. To overcome some of such variations, in the past we developed a bioinformatics pipeline named Proteogenomics Viewer, which was able to identify unknown alternative splicing events and quantify them at peptide level from different datasets using a unified peptide search engine approach followed by normalization of the data. We now further implemented the script to also perform quantitation at protein level, using the median distribution of all its identified peptides. We also tools for deep investigation of all datasets. In previous version of the Viewer, the user had to provide the gene of interest in order to load the data; now, the user can set quatitative parameters of interest, and the Viewer will retrieve all genes from a given sample which fulfill the requirements. For example, the user can setup a search looking for proteins highly expressed in an ovarian cancer dataset, and define how large is the expression level difference compared to the remaining datasets present in the Viewer. We foresee such analysis will be helpful to define molecular signatures in samples, as well as biomarker candidates in cancer cells. We provide some examples of such observations.

Possible Horizontal Gene Transfer in a Damage Suppressor Gene in tardigrades

16 Apr
1:00pm
P114

Jose Nelson Badziak Junior[1] and Edson Ricardo Junior[2]
[1]Independent Researcher; [2]UFRN

Tardigrada is a phylum of some of the most resistant animals that are alive. All of the known tardigrades have an unusual resistance to radiation, reactive oxygen species (ROS), temperature, desiccation and pressure among eukaryotes. Hashimoto et al. have shown a DNA-associated protein that suppresses DNA breaks caused by radiation and ROS in the Ramazzottius varieornatus, one species of tardigrade. Although no homology was found in the NCBI database (both for BLASTn and BLASTp) for this Damage suppressor (Dsup), one hypothetical protein was shown in another species of the tardigrade, the less radiation resistant Hypsibius dujardini, with similar profiles in hydrophobicity and charge distribution along the protein, which supposedly also functions in DNA protection against radiation and ROS. Nonetheless, the origin of this gene, so far unique in tardigrades, remains unclear. Some authors hypothesize that the Dsup gene might have arisen exclusively in tardigrada, under weak selective pressure during evolution. In this work we hypothesize that this gene might have arisen from horizontal gene transfer (HGT), most likely from bacteria. We quantified the relative GC content from this gene in both Ramazzottius varieornatus and Hypsibius dujardini compared to the CG content of the whole genome and of all genes, as a common method for analyzing HGT. We found that GC% is 8.79% higher in the Dsup gene compared to all genes and 18.45% higher compared to the whole genome of the Ramazzottius varieornatus (GC%: 56.28% in the Dsup gene, 51.73% in all genes and 47.51% of the whole genome). In the Hypsibius dujardini, the GC% of the Dsup hypothetical gene is 4.19% higher compared to the whole genome (GC%: 47.36% in the hypothetical Dsup gene and 45.45% in the whole genome). The higher GC content in the Dsup gene support the hypothesis of horizontal gene transfer as a mechanism for the origin of the damage suppression in tardigrades. If the hypothetical Dsup in Hypsibius dujardini is confirmed to be a damage suppressor similar to the Ramazzottius varieornatus' Dsup, the lower variation in Hypsibius dujardini's Dsup GC% indicates a weaker evolutionary pressure in this gene in this species compared to Ramazzottius varieornatus's Dsup. Furthermore, this would corroborate with the lower radiation resistance in Hypsibius dujardini due to mutations that carry the GC% to background level.

16 Apr
1:00pm
P115

Polymorphisms in plastid terminal oxidase (PTOX) from Arabidopsis ecotypes evidenced SNP-induced 3-D structure variants associated with altitude and pluviosity

Karine Thiers Leitão Lima[1], Geraldo Rodrigues Sartori[2], João Hermínio Martins da Silva[2] and José Hélio Costa[1]

[1]Fiocruz-CE; [2]UFC

Plant plastoquinol oxidase (PTOX) is a chloroplast oxidoreductase involved in carotenoid biosynthesis, chlororespiration, and response to environmental stresses. The present study aimed to gain insight of the potential role of nucleotide/amino acid changes linked to environmental adaptation in PTOX gene/protein from Arabidopsis thaliana accessions. SNPs in the single-copy PTOX gene were identified in 1190 accessions of Arabidopsis using the Columbia-0 PTOX as a reference. The identified SNPs were correlated with geographical distribution of the accessions according to altitude, climate, and rainfall. Among the 32 identified SNPs in the coding region of the PTOX gene, 16 of these were characterized as non-synonymous SNPs (in which an AA is altered). A higher incidence of AA changes occurred in the mature protein at positions 78 (31%), 81 (31.4%), and 323 (49.9%). Three-dimensional structure prediction indicated that the AA change at position 323 (D323N) leads to a PTOX structure with the most favorable interaction with the substrate plastoquinol, when compared with the reference PTOX structure (Columbia-0). Molecular docking analysis suggested that the most favorable D323N PTOX-plastoquinol interaction is due to a better enzyme-substrate binding affinity. The molecular dynamics revealed that plastoquinol should be more stable in complex with D323N PTOX, likely due a restraint mechanism in this structure that stabilize plastoquinol inside of the reaction center. The integrated analysis made from accession geographical distribution and PTOX SNPs indicated that AA changes in PTOX are related to altitude and rainfall, potentially due to an adaptive positive environmental selection.

Genomic landscape of genes and their non-coding RNAs

16 Apr
1:00pm
P116

Leandro Magalhães, Rafael Pompeu, Gilderlanio S. Araújo, Ândrea Ribeiro-dos-Santos and Amanda F. Vidal

Laboratory of Human and Medical Genetics, Institute of Biological Sciences, Federal University of Pará, Belém, Pará, Brazil.

Non-coding RNAs (ncRNAs) are a diverse class of transcripts produced by the eukaryotic genome and also represent the majority of RNAs produced by the cell. ncRNAs can have structural or regulatory activities, being piwi-interacting RNAs (piRNAs) and circular RNAs (circRNAs) examples of the latter. Some genes can produce both piRNAs and circRNAs but it is unknown what enables a gene to do so. We aimed to characterise the genomic features and context of the genes that most transcribe both piRNAs and circRNAs in order to understand what allows a gene to produce multiple types of ncRNAs. In order to achieve that, we downloaded: i) the list of all human piRNAs and their genomic origin from piRBase 2.0 [http://www.regulatoryrna.org/database/piRNA/index.html]; ii) the list of all human circRNAs and their coding genes from circBase [http://www.circbase.org] and iii) the list of all human genomic loci from Ensembl [http://www.ensembl.org - Release 95], then filtered the ones that do not code neither piRNAs nor circRNAs. We observed a total of 2515 genes that code both piRNAs and circRNAs, 1780 genomic regions that code only piRNAs, 9298 genes that code only circRNAs and 25683 genomic loci that does not code neither piRNAs nor circRNAs. We then selected 20 random genes that do not code piRNAs or circRNAs and the top 20 genes/genomic loci that: i) produce at least 10 piRNAs and 10 circRNAs; ii) produce only piRNAs; iii) produce only circRNAs. We annotated gene size, genomic location and orientation of the selected genes using GeneCards hg38 [genecards.org]; number of exons and introns using UCSC's Genome Browser hg38 [https://genome.ucsc.edu]; number of splicing variants using Ensembl; number of mRNA isoforms, unspliced forms and probable alternative promoters from AceView [https://www.ncbi.nlm.nih.gov/IEB/Research/Acembly/]; and number of transposable elements from Dfam 3.0 [dfam.org]. To compare the annotations between each group, we used an One-way ANOVA and Tukey's HSD test as post-hoc pairwise comparisons and considered p-values < 0.05 to be statistically significant. We found that gene size, number of introns, exons, splicing variants, mRNA isoforms, unspliced forms, probable alternative promoters and transposable elements were significantly different in the studied groups (ANOVA, $P \leq 6.9e\text{-}05$). We also observed that genes that produce both piRNAs and circRNAs have a higher number of splicing variants (average of 19.94), mRNA isoforms (average of 21.73) and probable alternative promoters (average of 7.05) than genes that did not produce any of those ncRNAs (average of 8, 8.68 and 2.75, respectively - Tukey's HSD, Padj = 0.00345; 0.00095 and 0.00424, respectively). For these very same annotations, there was no significant difference between genes that produced both piRNAs and circRNAs and genes that

produce only circRNAs (Tukey's HSD, Padj = 0.968; 0.325 and 0.153, respectively). When comparing genes that produce both ncRNAs and genomic loci that produce only piRNAs for those annotations, we only found the number of splicing variants to be significantly different (Tukey's HSD, average of 19.94 vs. 3.27, Padj = 3.8e-07). That could be explained due the lack of information regarding the genomic regions that code piRNAs and most of the annotations we selected were not available or not present in the databases. Our results show that the overall genomic features of a gene that produces circRNAs act in a dominant fashion upon the features of a gene that produces piRNAs. In that matter, for a gene to produce both piRNAs and circRNAs it has to have most of the characteristics of a gene that produces circRNAs. A lot of information about the ncRNAs genomic features are still unknown, specially about the genes that produce only piRNAs. Overall, our results demonstrate how complex is the ncRNAs genomic landscape and how incipient is our knowledge about it.

Funding Agency: CAPES (Biocomputacional-Protocol No. 3381/2013/ CAPES)

Metylathion profile in hepatic fibrosis: the regulatory role of heptapeptide angiotensin-(1-7).

16 Apr
1:00pm
P117

Caio Mateus da Silva[1], Hellen Kuasne[2] and Letícia Ferreira Ramos [1], Silvia Regina Rogatto[3], Karen C. M. Moraes[1]

[1]UNESP, Rio Claro; [2]McGill University; [3]University of Southern Denmark;

The liver is an important organ because it is responsible for various functions in the body. When suffering from changes from infectious, chemical or biological agents, the organ responds to these lesions with hepatic fibrosis. Hepatic fibrosis occurs due to the accumulation of extracellular matrix causing tissue healing. If the liver damage is continuous, the liver loses healthy tissue and has its function compromised. The primary cell responsible for this process is the hepatic stellate cell (HSC). They are characterized by a quiescent phenotype in healthy liver and the change to an activated phenotype when the liver receives lesions. Although the liver has a regenerative capacity, the options for the treatment are scarce, being the main treatment the removal of the aggressor stimulus. In this way, heptapeptide angiotensin-(1-7) [ang-(1-7)], a product of the renin-angiotensin system, known for its anti-inflammatory, anti-fibrotic, antithrombotic, antihypertensive, antiangiogenic properties and cardioprotective, which appears as an innovative therapy in the treatment of fibrosis. On the other hand, DNA methylation consists of the addition of methyl radicals to the gene structure, which contributes to the alteration of gene expression. Thus, the study of the DNA methylation pattern may help elucidate the mechanistic peculiarities that occur in cell transdifferentiation. In this context, evaluating the effects of ang-(1-7) on altering the DNA methylation pattern of activated HSC cells, to investigate its effects on reversal of the pro-fibrogenic profile is relevant. Thus, HSC cells of the LX-2 lineage, which are cells capable of transdifferentiating from a quiescent phenotype to a cell culture-activated phenotype, were used. To perform the analyzes, these cells were cultured in three different conditions: in the quiescent phenotype condition, the cells were cultured in medium containing 2% fetal bovine serum (FBS) and under the activated conditions the cells were cultured in medium containing 10% of FBS and when necessary, treated or not with 10-7 M ang-(1-7) (Merck - Millipore). After cultivation, the Infinium HumanMethylation450 BeadChip array (Illumina, Inc.) microarray platform was used to analyze the methylation profile of these cells on a large scale. The data obtained were processed using the R Core Team program (2014), in conjunction with the Bioconductor/wateRmelon data package. After statistical corrections and normalizations (BMIQ), differentially methylated probes were considered as those with $\Delta\beta$ <-0.10 or> +0.10. Subsequently, only differentially methylated, protein-bound probes were analyzed comparatively between cell groups. Functional annotation and gene ontology were performed with Blast2GO software and GeneOnthology.org. The analyzes demonstrated that 631 probes were differentially methylated in the transdifferentiation process of HSCs for their activated

phenotype. Among the activated and treated conditions with ang-(1-7), 289 methylated probes were found. However, when comparing the methylation profile of the quiescent cells to the treatment of the ang-(1-7) activated cells, a total of 1723 probes were hypermethylated and 24 probes were hypomethylated. In sequence, functional enrichment analyzes revealed 6 statistically relevant proteins in the cellular transdifferentiation process (SPX, GPER1, KCNB2, CHRM2, SRF, CHRM3). In addition, treatment with ang-(1-7) has shown that genes related to protein and ion binding mechanisms, migration and regulation of cell motility have been methylated. With this, it is concluded that with the presence of heptapeptide, there is an increase in the methylation levels with a higher representation of hypermethylated probes. In addition, the identification of gene sequences corresponding to differential methylation proteins responsible for cell transdifferentiation and proteins responsible for cell motility, which is related to the extracellular matrix, demonstrates an important regulatory role of heptapeptide in hepatic stellate cells and helps in understanding the biology of cellular activation.

Funding Agency: FAPESP (2013/21186-5)

In silico analysis points to prognostic value of EGFL7 in astrocytomas possibly related to angiogenesis through Notch pathway

16 Apr
1:00pm
P118

Bruno Henrique Bressan da Costa, Rui Manuel Reis and Lucas Tadeu Bidinotto
Molecular Oncology Research Center, Barretos Cancer Hospital

Introduction and objectives: EGFL7 is a protein normally secreted by endothelial cells during angiogenesis and vascular cords formation. Its expression is reduced in mature blood vessels and normal adult tissues. Several studies have been conducted to understand its potential role in tumorigenesis, since it has been found expressed in several tumor cells, and its role was found related to MAPK and Notch pathways, extensively studied in cancer. Our research group has previously found EGFL7 protein expression correlated to a poor prognosis in pilocytic astrocytomas. The objective of this work is to extend this analysis in gliomas using bioinformatic tools, and to investigate the prognostic value of EGFL7 in TCGA glioma data, as well as its mechanism of action. Material and methods: Log-transformed RSEM-normalized RNA sequencing data was obtained from TCGA Lower Grade Glioma (LGG; n=194 astrocytomas, 130 oligoastrocytomas and 191 oligodendrogliomas) and Glioblastoma (GBM, n=150) using TCGA2STAT R package. Patients were categorized into high or low EGFL7 expression according to the median, i.e., if the expression value of the gene was higher than the median, the expression was considered high, otherwise, it was considered low. Survival log rank analysis was performed in each tumor type considering EGFL7 expression. Additionally, in order to find genes co-expressed with EGFL7, Spearman's correlation of EGFL7 and the other 20,500 genes present in the RNA sequencing was performed. Those presenting correlation $> |0.6|$ were selected, and Enriched Gene Ontology (GO) analysis was performed. The statistical significance was considered when $P<0.05$. Results: The EGFL7 RNA expression was not changed according to the diagnosis (5.87+-0.55 for astrocytoma; 5.88+-0.57 for GBM; 6.00+-0.55 for oligoastrocytoma; 5.99+-0.61 for oligodendroglioma). Survival analyses showed that high EGFL7 expression was correlated to improved overall survival in grade II and grade III astrocytoma patients (5-year survival of 67.6% in high expression vs. 40.9% in low expression, P=0.003). EGFL7 expression was not correlated to differences in survival in GBM (1-year survival of 55.6% in high vs. 58.4% in low expression, P=0.373), oligoastrocytoma (5-year survival of 64.9% in high vs. 71.3% in low expression, P=0.743), and oligodendroglioma (5-year survival of 64.4% in high vs. 73.7% in low expression, P=0.529). Spearman's correlation showed 79 genes co-expressed with EGFL7 in GBM dataset, and 22 genes in LGG. Enrichment Gene Ontology analysis showed GO biological processes related to angiogenesis and lymphangiogenesis, vasculogenesis, endothelial cell differentiation and migration, extracellular matrix organization and regulation of Notch signaling pathway (FDR<0.05). Among the genes related to regulation of Notch signaling pathway, besides EGFL7, there were present NOTCH4 receptor and its ligands DLL4, JAG2, and

PEAR1, showing that this pathway may be regulated in tumors expressing high levels of EGFL7. Furthermore, the GO biological process "negative regulation of angiogenesis" is enriched in cells expressing high levels of EGFL7, with co-expression of ECSCR, TIE1, MMRN2, DLL4 and GPR4 genes. Conclusions: Patients diagnosed with grade II or III astrocytomas who presented high levels of EGFL7 may have improved survival possibly due to the co-expression of genes that regulate angiogenesis through Notch pathway. More studies are warranted in order to unveil the precise role of EGFL7 in this pathway.

Funding agency: Bruno Henrique Bressan da Costa is recipient of FAPESP fellowship (number 2018/20737-1) and Lucas Tadeu Bidinotto is recipient of FAPESP grant (number 2016/21727-4).

Machine learning approach to model breast cancer networks from heterogeneous data: a preliminary analysis on SEER data.

16 Apr
1:00pm
P119

Marcel da Câmara Ribeiro-Dantas, Nadir Sella, Vincent Cabeli, Honghao Li and Hervé Isambert

Institut Curie, PSL Research University, CNRS - UMR168, Sorbonne Université

We have developed a machine learning approach to analyze heterogeneous clinical data in order to assist health professionals and researchers to visualize direct and indirect interrelations that can lead to a better understanding of complex diseases. With the help of physicians from Curie Hospital specialized in breast cancer, we have applied this inference method of clinical networks to breast cancer which is the second leading cause of cancer death among women, and also the most diagnosed cancer in women. The reconstruction of such networks is illustrated here through the reconstruction of a breast cancer network based on data from the Surveillance, Epidemiology, and End Results (SEER) program. SEER has been collecting cancer incidence data since 1973 and currently contains data covering approximately 34.6% of the population of the United States. Among the different datasets it provides, there are over 500 variables with data ranging from demographics, medical records including biomolecular information, treatment data, socioeconomic data and county-level socioeconomic indexes. This global network analysis has been able to reconstruct not only associations that are already known in the literature, giving solidity to our method, but also uncovering unsuspected direct and indirect associations that are putative cause-effect relationships between clinically relevant information.

16 Apr
1:00pm
P120

Analysis and comparison of missense mutations in cancers with different etiologies and their structural consequences.

Marilia Viana Albuquerque de Almeida, Laise Cavalcanti Florentino, Danilo Lopes Martins, Jorge Estefano Santana de Souza, Sandro José de Souza and João Paulo Matos Santos Lima

UFRN

In Brazil, in the biennium 2018/2019 there will be about 1.2 million new cases of cancer according to the National Cancer Institute - INCA. Cancer is among the leading malignancies in the world, characterized by the disordered growth of cells, which invade tissues and organs. Omic technologies made it possible to obtain a deeper understanding of the genes related to cancer and their respective mutations: deleterious (MD) - related tumor progression; and neutral (MN) - unrelated to tumor progression. Computational tools and the exploration of Residue interaction networks (RINs) in proteins are being developed as ways of predicting these types of mutations. Recent studies also suggest that some types of cancer may be related to environmental and hereditary factors; and other types would be more strongly associated with stochastic factors such as random mutations that arise during DNA replication. The purpose of this study is to analyze and compare the missense (MM) mutations and their structural consequences related to stomach cancer (STAD), lung (LUNG) and glioblastoma (GBM). Genes and their respective mutations were mined from the Genome Atlas (TCGA) and Catalog of Somatic Mutations In Cancer (COSMIC) databases, and filtered by the Ndamage parameter (mutation impact classification). The MM that occurred in coding regions were selected. Three-dimensional protein structures (PDBs) were obtained from the pdb database using UniProt codes. The mapping of the position of the PDBs was carried out by the SIFTS database and the structural effect of the mutations was estimated based on the construction of RINs, using RING 2.0 software. Files containing Nodes (each amino acid of the protein) and Edges (connections between amino acids in the form of chemical interactions) were generated, such information added to the database and related to the mutations. The data were stored in MySQL. The Degree (connectivity between Nodes) was observed and this parameter was limited to 15. It was noted that for the types of cancer in this study that there is a high amount of mutations in Degrees with a value of up to five. As the Degree increases, the amount of MN decreases and the ratio of MD increases. In STAD and GBM, there is an increasing tendency of MDs as Degree increases; in LUNG the MDs remain constant and after increasing Degree to 11, an increase of this type of mutation is noticed. In relation to the amino acid exchange (AA), there is a higher proportion in STAD between AA of polar type for aromatic and non-polar for positive charge; in GBM, from aromatic to positive charge and positive charge to positive charge, and in LUNG from aromatic to polar and non-polar to positive charge. Regarding the Edges, in STAD and GBM,

there was a higher proportion of MDs for dislocated and pication-type interactions. And in MN, ionic and pipstack. In LUNG, in the MD, a higher proportion occurred for interactions of type IAC and hydrogen bonds. And in MN, disulfide and pipstack bridges. Mutations are in low degrees, but MDs tend to occur in nodes that are in higher degrees, as it is expected that greater connectivity may cause greater impact on proteins. Alterations in chemical groups and their interactions may be associated with possible structural changes in proteins, with potential to affect biological activity. From the evaluation of its impacts on the RIN, one can predict the effect of a mutation and associate it with the onset and progression of these types of cancer.

16 Apr
1:00pm
P121

A pan-cancer analysis of chr9p22.1-21.3 locus unveils genes potentially important in the development of several tumor types

Paola Gyuliane Gonçalves[1], Rui Manoel Reis[1] and Lucas Tadeu Bidinotto[2]
[1]Barretos Cancer Hospital; [2]Barretos School of Health Sciences

Introduction and objectives: There is a crucial demand to identify molecular markers for cancer to be able to improve the personalized treatment, diagnosis and prognosis in several types of cancer. Our research group found that 50% of glioblastoma patients present loss of heterozygosis (LOH) of the 22.1-21.3 region of the short arm of chromosome 9. The loss of this locus has been linked to the development of several types of cancer. Therefore, the aim of this study was to identify, using in silico tools, the role of the genes present in the locus chr9p22.1-21.3 in 31 types of cancer from The Cancer Genome Atlas (TCGA), as well to associate this data with clinical-pathological features and describe potentially new driver genes with clinical impact in the carcinogenesis. Material and methods: CGH microarray (aCGH) data of chr9p22.1-21.3 region was imported from 10,392 TCGA samples across 31 tumor types in R using TCGA2STAT package. The patients were considered with region loss if the value in any gene of the region was lower than -0.1. Overall survival was analyzed using log rank statistical analysis. From the datasets with statistically significant differences between "loss" and "no loss" in aCGH, RSEM-normalized RNA sequencing data was obtained from the 24 genes present in the locus of interest. For each gene/dataset, the patients were categorized into high expression and low expression based on the median expression for dataset, i.e., if the expression value of the gene was higher than the median of the dataset, the expression was considered high, otherwise, it was considered low. The differences of the groups with $P<0.05$ were considered statistically significant. Results: From the 31 aCGH datasets analyzed, twelve presented more than 50% of the patients with loss in chr9p22.1-21.3 locus: adrenocortical carcinoma (ACC), bladder urothelial carcinoma (BLCA), cholangiocarcinoma (CHOL), esophageal carcinoma (ESCA), glioblastoma (GBM), head and neck squamous cell carcinoma (HNSC), lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), mesothelioma (MESO), pancreatic adenocarcinoma (PAAD), skin cutaneous melanoma (SKCM) and uterine carcinosarcoma (UCS). Loss in chr9p22.1-21.3 was associated to a poorer prognosis in eleven datasets ($P<0.05$): GBM, HNSC, MESO, PAAD, kidney clear cell carcinoma (KIRC), kidney papillary cell carcinoma (KIRP), low grade glioma (LGG), lung adenocarcinoma (LUAD), pheochromocytoma and paraganglioma (PCPG), sarcoma (SARC) and uterine corpus endometrial carcinoma (UCEC). Considering mRNA expression, eleven genes presented relevant expression values (SLC24A2, MLLT3, KIAA1797, PTPLAD2, KLHL9, MTAP, C9orf53, CDKN2A, CDKN2B, DMRTA1 and ELAVL2). For GBM, the patients whose gene DMRTA1 was highly expressed had a worse survival

when compared with the patients that had low expression (p=0.04), while in LGG the high expression of the genes MLLT3 (p=2.05E-07), KLH9 (p=2.80E-05), MTAP (0.0006) and ELAVL2 (0.0002) are associated with a better prognosis and PTPLAD2 (p=1.37E-05) with a poorer prognosis. For KIRC, high expression of MLTT3 (p=0.035), KIAA1797 (p=0.024), KLH9 (p=1.45E-08), MTAP (p=0.01), CDKN2B (p=0.007), DMRTA1 (p=1.77E-07) and ELAVL2 (p=0.006) were associated with a better survival and CDKN2A (p=0.005) were associated to a worse survival. The high expression of CDKN2A was also associated with a worse survival in PCPG (p=0.017) and UCEC (0.0006). In the KIRP dataset, the high expression of CDKN2B (p=0.007) was associated with a poorer prognosis. In LUAD the high expression of the genes MLLT3 (p=0.023), KLHL9 (p=0.032), PTPLAD2 (p=0.003) were associated with better prognostic, as well the genes SLC24A2 (0.043), KIAA1797 (0.0003), PTPLAD2 (0.009), KLHL9 (1.55E-07), CDKN2B (0.0001), DMRTA1 (0.028), ELAVL2 (0.007) in MESO. While the high expression in the genes ELAVL2 (0.011) was correlated to a better prognosis in PAAD, PTPLAD2 (0.0013) in SARC and KIAA1797 (0.004) in UCEC. Conclusions: Almost all the genes analyzed in kidney clear cell carcinoma (except for SLC24A2 and PTPLAD2) and mesothelioma (except for MLLT3), presented any prognostic importance, showing that, possibly, it is not the function of the gene per se that is impacting the tumor development and/or progression in these tumors, but the loss of the locus, as shown in the analysis of CGH microarray. Moreover, besides CDKN2A and CDKN2B, there are several other genes in chr9p22.1-21.3 that are possibly related to cancer development. They may be related to cell cycle, differentiation, and metabolism that might be important of future studies, specially in glioma, lung adenocarcinoma, kidney renal clear cell carcinoma and mesothelioma. Furthermore, the results of these studies may contribute to the development of new therapies targeting specific molecules in those cancers. Funding Agency: PGG is recipient of FAPESP fellowship (FAPESP number 2017/09749-5) and LTB is recipient of FAPESP grant (2016/21727-4).

16 Apr
1:00pm
P122

Non-coding RNAs networks evidence the epigenomic regulation complexity

Rafael Pompeu, Leandro Magalhães, Ândrea Ribeiro-dos-Santos and Gilderlanio S. Araújo, Amanda F. Vidal

Universidade Federal do Pará

Non-coding RNAs (ncRNAs) represent more than 70% of human genome and have broad structural, functional and regulatory roles. Several functional studies have suggested them as promising biomarkers and therapeutic targets for many diseases, but most of those ncRNAs do not have their role fully understood. We constructed a ncRNA-gene network that includes three classes of ncRNAs − microRNAs (miRNAs), piwi-interacting RNAs (piRNAs) and circular RNAs (circRNAs) with either their target (for miRNAs) or coding genes (for piRNAs and circRNAs). To achieve that, we used three databases to extract these ncRNAs-gene interactions: i) miRTarBase 7.0 [http://mirtarbase.mbc.nctu.edu.tw/php/index.php], that provided the experimentally validated human miRNAs-target gene interactions; ii) piRBase 2.0 [http://www.regulatoryrna.org/datab which provided the corresponding genomic origin of all human piRNAs; and iii) circBase [http://www.circbase.org], that provided the host genes of all human circRNAs. We also performed functional analysis using STRING v11.0 (https://string-db.org) to predict the biological processes and KEGG pathways of the identified genes. These datasets have a total of 648 miRNAs targeting 2338 genes, 7948 piRNAs being coded by 4295 genes and 91032 circRNAs being coded by 11813 genes. We found 6708, 21277 and 91032 interactions between these genes and miRNAs, piRNAs and circRNAs, respectively. In miRNA-target gene network, hsa-miR-155-5p presented the highest number of connections, targeting 137 genes. Functional analysis of these 137 genes showed enrichment for regulation of developmental process and for pathways in cancer, specially in colorectal cancer. We also performed functional analysis of the ten genes with the highest number of miRNAs-connections (PTEN [55], CDKN1A [39], BCL2 [39], CCND1 [38], IGF1R [33], CDK6 [31], MYC [28], STAT3 [27], HHMGA2 [25] and VEGF [24]) and found enrichment for positive regulation of cell population proliferation and pathway of miRNAs in cancer. In the piRNA-gene network, piR-hsa-2100 showed the highest number of connections, being produced by 581 different genes. Although this piRNA seems to have a great relevance by being produced by several genes, its biological functions remain unknown and need to be further investigated. It is interesting to note that the two genes with the highest number of piRNAs-connections (FAM225A and FAM225B) − coding 639 and 630 piRNAs, respectively − belong to the same family of intergenic non-coding RNA genes and share most of the produced piRNAs (n = 628). These two genes have their function unknown. In the circRNA-gene network, all circRNAs have only one connection, which confirms that a same circRNA cannot be produced by different genes. Furthermore, we found that BIRC6 is the gene

that originated the highest number of circRNAs isoforms (444 circRNAs-connections). This gene (261.872 bp) can produce only 32 mRNA isoforms, suggesting a much higher preference of the non-canonical splicing over the canonical splicing. BIRC6 is related to the apoptotic process and with several diseases, such as small cell lung cancer and hepatitis B. Overall, our results showed that the ncRNAs networks have many connections and may reflect how complex and dynamic the epigenetic mechanisms are. This study showed that ncRNAs and its related genes are associated with important biological processes and with several human diseases, specially cancer, indicating that these networks are a potential tool for predicting interactions, measuring clinical-related impacts and searching for novel biomarkers and therapeutic targets.

16 Apr
1:00pm
P123

Noisy Neighbors Algorithm: Concept and Implementation in Python

Suzana Pôrto Almeida[1], Rafael Mendonça Colares[1], Francisco Eder de Moura Lopes[1], Raul Victor Medeiros da Nóbrega[2], Saul Gaudencio Neto[1], Leonardo Tondello Martins[1], Ana Cristina de Oliveira Monteiro Moreira[1], Antonio Carlos da Silva Barros[2], Pedro Pedrosa Rebouças[2], Victor Hugo Costa de Albuquerque[3] and Kaio César Simiano Tavares[1]

[1]Laboratório de Biologia Molecular e do Desenvolvimento, Universidade de Fortaleza, Fortaleza, Ceará, Brazil; [2]Laboratório de Processamento de Imagens e Simulação Computacional, Instituto Federal de Ciência e Tecnologia do Ceará, Fortaleza, Ceará, Brazil; [2]Programa de Pós-Graduação em Informática Aplicada, Universidade de Fortaleza, Ceará, Brazil

INTRODUCTION AND OBJECTIVES Genome Wide Association Studies (GWAS) are studies conducted with a great amount of individuals to identify relevant SNPs (single nucleotide polymorphisms) for specific traits. Even though most GWAS focus on gene analysis, complex traits might be associated with non-coding DNA, which contains cis-regulatory elements, such as enhancers and promoters. On the other hand, cis-elements depend on trans regulators, known as transcription factors, binding in specific sites to be able to effectively operate in gene expression regulation networks. All of this regulatory mechanism orchestra compose an emergent field highly significant to systems biology called Regulomics. Although new evidences show that the inheritage of SNPs in haplotypes can explain better the contribution of genetic factors in complex traits, up until now, this kind of approach was not used to evaluate the presence of SNPs inside cis-regulatory elements. Based on this statement, the aim of this work was to develop an algorithm to process SNP haplotypes and make DNA sequences, if very close SNPs are detected within these haplotypes, so it could be used in software to analyze SNPs alleles combinatorial effects in transcription factor binding.

MATERIAL AND METHODS Using Python through Google Colaboratory, we applied object oriented paradigm for a better data modeling process and the creation of a data structure based in graphs. The initial part of the algorithm was developed to process haplotypes provided by the user, from GWAS based studies in haplotypes, as well as search in dbSNP for information about genomic location, major and minor alleles for each SNP. If the inserted SNPs are close enough, less than 32 nucleotides of distance between each other, a DNA sequence block will be made, the SNPs will stay in the same block and every possible combination of alleles in these SNPs will be computed in different DNA sequence blocks. If a block has n SNPs and a(i) is the number of possible alleles in i SNP, the number of sequences to be analyzed is based on this product equation: a(1)*a(2)*a(3)*...*a(n).

RESULTS AND DISCUSSION The algorithm creates all these DNA sequences in FASTA format and stores it in a file with .fna extension, for each sequence a header

is made to identify its wild-type allele and variation allele. The generated file and its individual sequences can be used in a software that analyses transcription factors. The name given to our algorithm was inspired in a simple analogy. Considering the Human Genome as a very large city containing trillions of houses (nucleotides). The noises (mutations) in each residence is an aspect that sometimes can be undesirable (complex traits). Until now, researchers were more interested in evaluate the effects of noise pollution only in more wealthy neighborhoods (genes), forgetting suburbs (non-coding regions) streets overlapping noise (SNP-haplotype). With this work, we would like to address the importance of these regions too.

CONCLUSIONS Therefore, our algorithm was developed to, not only investigate noises coming from individual houses (Single-SNP), but the noise from each forgotten suburban street (SNP-Haplotype). And then, with all of this, we intend to reach for a holistic vision of what is happening in the city.

REFERENCE KHANKHANIAN, P. et al., Haplotype-based approach to known MS-associated regions increases the amount of explained risk. Journal of Medical Genetics, v. 52, n.9, p. 587-594. 2015.

Funding Agency: CNPq

16 Apr
1:00pm
P124

Information Theory and scTranscriptomics: selection of relevant transcripts in single-cell datasets by FCBF

Tiago Lubiana Alves and Helder Takashi Imoto Nakaya
University of São Paulo

Single cell transcriptomic (scTranscriptomics) data analysis implies sifting through datasets containing thousands of cells and tens of thousands of features. Selecting the variables (transcripts) that hold relevant information about a characteristic of the cells (e.g., cell types) is a crucial first step for downstream processing. The detection of essential genes is traditionally done with differential expression (DE) methods designed only for expression data, such as edgeR and DESeq2. Information theory is one a field of research with validated applications for feature selection outside the biomedical field, but, to date, no study has systematically applied information-theory based algorithms to scTranscriptomics. To explore this area, we developed an R/Bioconductor package for binarization (on/off) of expression and redundancy-minimizing feature selection based on conditional entropy of transcript levels and cell class (FCBF, Fast Correlation-Based Filter by Yu and Liu, 2003)). For validation, we applied our processing pipeline to a Peripheral Blood Mononuclear Cell single-cell dataset and were able to detect genes known to be important for immune cell function. Moreover, some FCBF-detected genes were not highlighted by standard DE approaches, demonstrating the potential of the technique for novel discoveries in scTranscriptomics.

A stochastic spatial model for heterogeneity in cancer growth

17 Apr
1:00pm
P201

Alexandre Sarmento Queiroga[1], Mauro César Cafundó Morais[1], Tharcisio Citrangulo Tortelli Jr[2] and Roger Chammas[1]

[1]USP; [2]ICESP

Establishing a quantitative understanding of tumor heterogeneity, a major feature arising from the evolutionary processes taking place within the tumor microenvironment is an important challenge for cancer biologists. Recently established experimental techniques enabled a summary of the variety of phenotypes exhibited by tumor cells by classifying them as proliferative or migratory. In the former, cells mostly proliferate and rarely migrate, while the opposite happens with cells having the latter phenotype, a "go-and-grow" description of heterogeneity. In this manuscript, we present a discrete time Markov chain to simulate the spatial evolution of a tumor which heterogeneity is described by cells having those two phenotypes. The cell density curves have two qualitatively distinct regimes. One recovers the Gompertz curve widely used for tumor growth description. The second is a bi-phasic growth which temporal shape resembles the tumor growth dynamics under the influence of immunoediting. We also show how our representation of heterogeneity give rise to variable spatial patterning even when the tumors have similar size and dynamics.

17 Apr
1:00pm
P202

Origin of the prostanoid system that participates on the control of menstrual cycle.

Andre Luiz Garcia de Oliveira[2], Fenícia Brito Santos[2], Tetsu Sakamoto[1] and José Miguel Ortega[2]

[1]UFRN; [2]Universidade Federal de Minas Gerais

Menstruating species are higher order primates, including humans and Old World monkeys. The proliferative phase of the menstrual cycle is characterized by endometrial re-growth consisting of endometrial cell proliferation and angiogenesis. Prostanoid system is associated with specific stages of the menstrual cycle and expression predominantly increases from the mid-secretory phase through to, and including, menstruation. It has been reported a comprehensive view of prostanoid action in the endometrium during the menstrual cycle. We thus set to analyze the clade of origin of the main components of the prostanoid system along evolution. The methodology relied on searching for orthologues taxonomic distribution with the software produced by our group TaxOnTree and depicting the clade of origin of each component. Then a chart was drawn with PathVisio to allow for computational coloring according the clade of origin. We observed that: (1) During menstruation there is prevalence of PGF2alpha and PGE2, which involves the activity of four enzymes. They are originated by Euteleostomi (PTGS2 and PTGE3), Theria (AKR1B1) and Eutheria (CBR1). Therefore menstruation is a relatively new addition. During the subsequent phase, endometrium proliferation (2), there is a remarkable expression of PTGFR receptor, originated by Euteleostome. Secretory phase (3) is then divided in Early (ES), Middle (MS) and Late Secretory (LS) phases: ES (4) is controlled by expression of AKR1C3, product of recent duplications involving AKR1C1 to C4, but the full branch is originated by Eutheria. In this phase there is also expression of HPGD (Eukaryota) and PTGER1 receptor, originated by Euteleostomi. MS phase (5), under predominance of PGD2 regulation, has expression of PTGS1 and PTGES2, both originated in Euteleostomi, PTGDS (Mammalia) and TBXAS1, more ancient, originated in Bilateria. This period is marked by expression of several receptors, all originated by Euteleostomi: PTGER2, PTGER3, PTGER4 and TBXA2R. Finally, the period LS (6), controlled by the action of TXA2/PGI2, shows expression of PTGIS (Gnathostomata), PTGES (Mammalia), HPGDS (Euteleostomi), and the following receptors: PTGDR, PTGIR and TBXA2R (all from Euteleostomi). In conclusion, an important regulation of endometrium development accessed by prostanoid system relies on genes that appeared in the ancestor of man and fish (Euteleostomi clade), with some key contributions in clades Theria and Eutheria. Therefore the recent adaptation through menstruation relies on recruitments of pre-existing components in respect to the involvement of prostanoid system.

MiRNA Differential Expression in Tuberculosis

17 Apr
1:00pm
P203

Arthur R. Santos[1], Wanderson Gonçalves[2], Rafael Pompeu[1], Cleonardo Silva[2], Ândrea Ribeiro-dos-Santos[3], Sidney Santos[3] and Gilderlanio S. Araújo[4]

[1]Institute of Technology, Federal University of Pará, Belém, Pará, Brazil Laboratory of Human and Medical Genetics, Institute of Biological Sciences, Federal University of Pará, Belém, Pará, Brazil; [2]Laboratory of Human and Medical Genetics, Institute of Biological Sciences, Federal University of Pará, Belém, Pará, Brazil Center of Oncology Research; [3]Laboratory of Human and Medical Genetics, Institute of Biological Sciences, Federal University of Pará, Belém, Pará, Brazil Center of Oncology Research, Postgraduate Program of Genetics and Molecular Biology, Institute of Biological Sciences, Federal University of Pará, Belém, Pará, Brazil; [4]Laboratory of Human and Medical Genetics, Institute of Biological Sciences, Federal University of Pará, Belém, Pará, Brazil; Postgraduate Program of Genetics and Molecular Biology, Institute of Biological Sciences, Federal University of Pará, Belém, Pará, Brazil.

Tuberculosis, a disease caused by Mycobacterium tuberculosis infection, is the leading cause of mortality from a single infectious agent and is among the top 10 causes of death worldwide. A third of the world population carries latent mycobacterium and as consequence the potential evolution to active tuberculosis.The identification of individuals with high risk of developing active tuberculosis would allow strategies for a more effective prophylactic treatment and would prevent the evolution of the disease to highly infectious symptomatic stages, avoiding its transmission. One of the essential mechanisms for maintenance and healthy functionality of the cell is gene regulation. Changes in gene expression can have detrimental impacts on the cellular state and even lead to disease. MiRNAs may regulate the host response of tuberculosis, being a critical mechanism for the establishment of the infection, however, the dynamics of miRNAs expression and the implications for immune response in the lungs are unknown. Few studies on the expression profile of miRNAs in tuberculosis have been published. This technology offers a higher accuracy and sensitivity identification of miRNA sequences. Therefore, the objective of the present aims to evaluate the global miRNA expression profile (miRnome) of patients with active tuberculosis and their respective physicians without tuberculosis and to verify potential biomarkers of this condition. For this purpose, 22 whole blood samples were collected using Tempus Blood RNA tube, the total RNA was extracted using MagMAX Isolation Kit and quantified with NanoDrop-1000 spectrophotometer, these samples were divided in 15 tuberculosis patients (group 1) and 7 control samples (group 2). Data obtained from the sequencing process (Performed in Illumina MiSeq System) was submitted to a quality control pipeline using Trimmomatic program, aligned with the human genome using STAR software and quantified using HTSeq package. The differential analysis processes were performed using EdgeR program package. MiRNAs with adjusted p-value < 0.05 and $|\log2\text{FoldChange}| > 1$ were isolated and considered as differentially expressed (DE). A total of 210 expressed

miRNAs (total raw read counts > 10 in all samples) were identified in this analysis, which 148 were considered DE, among these miRNAs, has-miR-486-5p had the highest expression in both groups, followed, in order, by hsa-miR-92a-3p and hsa-miR-181a-5p in group 1 and hsa-miR-16-5p and hsa-miR-92a-3p in group 2. To narrow the number of target genes, 114 out of the 148 DE miRNAs were divided in 2 new groups using a threshold of $|\log2\text{FoldChange}| > 2$. The down-expressed group composed of 83 DE miRNAs, which 59 presents experimental interaction validation with 1058 genes, while the up-expressed had 31 DE miRNAs, 21 shows validated interaction with 526 genes. The identified genes are enriched for inflammation mediated by chemokine and cytokine signaling pathway, T and B cell activation and apoptosis signaling pathway.

GENOME-WIDE COMPARISON OF P-ATPASE FAMILY FROM EUKARYA

17 Apr
1:00pm
P204

Clara Dáfne Alves de Farias, Demitry Messias dos Santos, Osman Cavalcante Junior and Leonardo Broetto

Universidade Federal de Alagoas

P-type ATPases, or E1-E2 ATPases, are multiprotein enzymatic complexes known to be large ion pumps widely distributed in the life-conserving and highly conserved domains found in energy-transducing membranes of bacteria, chloroplasts and mitochondria in the Archaea, Bacteria and Eukarya domains. P-ATPases have three cytoplasmic domains (A, N and P) and two transmembrane domains (T and S) formed by several helices. In addition, P-ATPases have two conformations (E1 and E2) that interconvert from the binding of ATP and transfer of phosphorylase, resulting in the energy release during a phosphorylation and dephosphorylation cycle, this occurs through membranes against an electrochemical gradient. In eukaryotes, many families of P- ATPases have not yet been identified, the predominant ones in eukaryotes pump calcium in the membranes (P2A and P2B) and sodium (P2D), regulate intracellular pH like the proton pump (P3A), pump magnesium (P3B), carry phospholipids (P4) and regulate endoplasmic reticulum (P5A) homeostasis and are present in the lysosomal membranes of animals (P5B). The latter, P5- ATPases have been found only in eukaryotes and are still poorly understood, but mutations in this family have been found to be related to the variety of neurological diseases. Thus, the main objective of this work is to prospect the class of ATPases proteins in predicted genomes of organisms representative of the Eukarya domain, through in silico analyzes. A collection of genomes was generated from the genome database of NCBI (National Center for Biotechnology Information), then the annotation pipeline was performed from a procedure that consisted of two steps of annotation and identification. The identification step was performed using the HMMER 3.2.1 program package, with the hmmsearch program of models from the hidden Markov model derived from Pfam's flatfile alignment. The annotated proteins were compared using the Blastp (Stand-alone BLAST application - BLAST + 2.8.1) program against the NCBI non-redundant protein database. The alignment of 1477 prospected proteins was used to obtain phylogenetic inference from the Fasttree software and the interactive tree of life (iTOL) online tool to model the phylogenetic tree. From the analysis of 25 predicted genomes of model organisms widely distributed in the kingdoms: Protista, Fungi, Plantae and Animalia, belonging to the Eukarya domain, the consensus of the conserved domains was obtained with the alignment of these same 1,477 prospected proteins. With the distribution of P-ATPases observed by phylogenetic inference, it was possible to perceive broad representation in all the genomes of the collected species. The species Homo sapiens and Drosophila melanogaster, both from the Animalia kingdom, were more represented in the tree with 319 and 105 replications

in the terminal taxa, respectively. In H. sapiens, proteins with phospholipid transport function (P4 type-flipase pumps) and calcium transport in the plasma membrane (P2B-type ATPase-calcium pumps) were the most repeated. In the distribution table, the functions of 2,019 proteins were analyzed, with 48 related to the heavy metal bombs (P1 type ATPase) in the species Selaginella moellendorffii, in Physcomitrella patens 32 proteins are P2 type ATPases. In total, there were 1,446 identified and 573 hypothetical unidentified because they did not have annotations of the functional regions in the database or because doubts about their functions arise. Preliminarily, based on the data generated in this work, we verified that the phylogeny of the P-ATPases is highly conserved and consistent with the divergence of the genomes of the analyzed taxa, demonstrating its potential as molecular clock in the reconstruction of the ancestry of the eukarya domain. Analysis of molecular aspects of the family may provide clues to the expansion of early ancestor organisms in the domain, as well as help us understand the current distribution of P-ATPases in Eukarya and, possibly, the adaptive evolutions related to them in these genomes.

Prediction of pre-eclampsia and eclampsia with machine learning

17 Apr
1:00pm
P205

Cristina Wide Pissetti1[1], Rebeca Andrade Bivar[2], Sarah Cristina Sato Vaz Tanaka[3], Sueli Riul Silva[4] and Marly Aparecida Spadotto Balarin[3]

[1]Centro de Ciências Médicas, Universidade Federal da Paraíba, Brazil; [2]Centro de Informática, Universidade Federal da Paraíba, Brazil; [3]Instituto de Ciências Biológicas e da Natureza, Universidade Federal do Triângulo Mineiro; [4]Programa de Pós Graduação em Atenção à Saúde, Universidade Federal do Triângulo Mineiro

Preeclampsia is a pregnancy associated disease. It is characterized by high blood pressure and symptoms that are indicative of damage to other organ systems, most often involving the liver and kidneys. If left untreated, the condition could be fatal to mother and baby. This study aimed to predict the diagnosis of preeclampsia and eclampsia by analyzing a database containing clinical and genetic information about a group of 218 women, using supervised machine learning methods - Naive Bayes and Support Machine Vector (Support Vectors Machine - SVM) with its default parameters, executed in the Weka Version 3.8 software. The database contained 218 instances and 11 entry attributes (smoker or not, number of pregnancies, number of miscarriages, multiple pregnancy or not, blood pressure, familial recurrence, proteinuria level and 4 genetic variations of 3 genes (CASP8, FAS and IL-10) and a target attribute, which is the presence or absence of syndromes. All the responses given by pregnant women that were limited to "yes" and "no" were respectively transformed into 1 and 0. In addition, 3 variations in the 3 genes studied (CASP8 A/G and A / T), FAS (A/G) and IL-10 (A/G) had their nominal values (nucleotide variations) converted to numerical values, as well as blood pressure variation (hypotension, normal blood pressure, mild hypertension, moderate hypertension, severe hypertension, systolic hypertension and when missing data), converted to values from 0 to 6. The other attributes were normalized between 0 and 1.The validation method used was the Cross Validation with 10 groups. With the large number of missing data, the learning processes were divided into executing the methods even with the missing data (considering a numerical value representing that category, when it occurred) and executing after the database was cleaned - resulted in a base with only 37 instances. With the same databases, the balancing method, Spread Subsample, was used to reduce the number of instances of the majority class to a number equal to the quantity of the minority class. With the Naive Bayes method with unbalanced data, we obtained the accuracy of: 86.34% (with missing data) and 83.78% (without missing data). With balanced data using Spread Subsample: 85.45% (with missing data) and 53.84% (no missing data). With the SVM method and the same experiments, respectively, we obtained: 89.97%, 78.37%, 83.63% and 65.38%. Thus, the potential of the method for the diagnosis of diseases is observed, however, for the unbalanced set of attributes (without missing data), even with high acuities (83.78% and 78.37%), 30% and ~ 50%, respectively, of minority class cases (normal patients) were

misclassified. In this way, it is intended to increase the number of patients without the syndromes studied here, thus reducing the missing data and increasing the number of instances of this class, here minority, balancing the total set. In addition, we can apply methods that relate a different weight to the accuracy of the studied classes, generating better models. With balanced data, we had a reduced number of instances, 37, which resulted in a maximum of 65.38% accuracy in the database without missing data. The presence of missing data generates a category of non-existent values, which can lead to the creation of an unreal model, given the nature of the problem, since attributes that may belong to other categories are being treated as the same. Therefore, larger studies, considering the presence of missing data should be performed. Funding Agency: CNPq (446914/2014-2)

Using networks to ease the biological interpretation of Gene Ontology enrichment analysis results

17 Apr
1:00pm
P206

Diego Arthur de Azevedo Morais and Rodrigo Juliani Siqueira Dalmolin
Universidade Federal do Rio Grande do Norte

A Gene Ontology (GO) enrichment analysis produces, as output, a list of several terms related to a given ontology (Biological process, Molecular function, or Cellular component). Due to the hierarchical organization of the Gene Ontology, some of these terms may be involved in a parent-child relation. Therefore, a list containing dozens of terms, some of them related to a same less specific term, makes the biological interpretation a laborious task. To ease this task, it is common to use the Gene Ontology hierarchical information to create a similarity matrix containing the similarity ratio, a value between 0-1, between each couple of terms. Thus, using the information provided by this matrix, it is possible to generate tile plots, dendograms, hierarchical networks, or even remove redundant terms. This work uses the transcriptogramer R/Bioconductor package enrichment analysis output to generate a hierarchical network. The enriched GO terms from a given cluster were used to generate a network from a dendogram of biological processes. In this representation, enriched GO terms are depicted as nodes, and the distance between nodes indicates whether the terms are related to the same genes. Briefly, the distance is shorter when the Jaccard's Index calculated on the genes of two terms is close to 1. The node size represents the number of genes related to a given term, and the node color represents the rate between the number of genes considered significant in the differential expression, and the sum of all annotated genes of that GO term. Finally, the node color is normalized by the max rate found for the terms in a cluster. The result is a network containing groups of terms close to each other when they belong to the same mother biological process, easing the biological interpretation and the extraction of biological insights from the results of an enrichment analysis.

17 Apr
1:00pm
P207

Means of response to virus database: MRV-db

Elisson N. Lopes[1], Carlos A. X. Gonalves[1], Lissur A. Orsine[1], Rodrigo J. S. Dalmolin[2] and J. Miguel Ortega[1]

[1]Laboratorio de Biodados, Instituto de Ciencias Biolgicas, UFMG; [2]Departamento de Bioquimica, Centro de Biociencias, UFRN

Many people has become frustrated after selecting genes differentially expressed, mostly if with microarray technology, and when inspecting the results in a gene by gene bases realized that the expression levels of a couple of controls were mixed with the levels for threated samples. Moreover, some deceptions grown while looking in systems biology databases the scenario of action of the differentially expressed genes, because very often these genes are not found in system biology databases, for example, no pathway is found in Kegg or Reactome. Aiming to study host genes implicated in the response to viral infections we set up an approach based on a platform which provides high reproducibility, the Illumina expression beadchip. In this platform, a given gene is represented around 30 times; therefore any measurement provides 30 replicates for any gene. We searched for the platform most used with deposited series and choose the Illumina HumanHT-12 V4.0 expression beadchip. The platform identified in GEO as GPL10558covers 25,440 genes represented by over 47 thousand probes. Is has been used with several samplings, respectively: seven times for HIV-1, seven for H1N1, three for HCV, two for HRSV, two for SIV, two for HSV-1 and once foreach of these viruses: CMV, HERVH, LRTI, VSV and WNV. We processed locally the data using local Perl scripts and, by using Perl database independent interface, controlled a local MariaDBmysql database. To model the procedure we selected a GEO series on H1N1 infection (GSE29385Influenzavirus serotype association to global whole blood transcriptional changes), which comprises several different infections, being nine patients infected with H1N1 and provides 75 samples of control individuals. First, by using an online GEO implementation of differential analysis, GEO2R, we selected genes that presented the thresholds of <= 0.001 for p-value and <= 0.05 for adjusted p-value. The selection resulted in, respectively, 1313, 10082, and three entries for each of the three time points of infection: (a) Less than 72 hour after symptoms; (b) 3 to 7 days after symptoms and (c) 2 to 5 weeks after symptoms. With the objective of focusing on genes that would resist to a gene-by-gene inspection, we further selected the entries with the criteria that the total of infected samples would cluster together after clustering with K-means set for two clusters. The resultant genes were graded as gold standard in the database, and for the three experimental times they were, respectively: 654, 1331 and zero genes. We then analyzed the frequencies of presence of control samples clustered with the nine infected samples and to better select for differentially expressed genes we repeated the GEO2R analysis with a balanced number of controls, by using the control RNA samples that least clustered with infected samples. Therefore the resultant number of entries

was, respectively: 2359,282, and zero entries. From these, in 1282, 128, and zero cases que k-means clusters comprised all nine infected samples, graded as gold, as opposite to silver, denoting genes that have just passed the statistical test. Thus, the phenomena in which the gold standard genes were involved might represent system biology scenarios that really means to the viral response, and therefore representing the means that a cell uses to response to viral infection. They were used to support the Means of Response to Virus database, MRV-db. Clearly both uses of the word means relate to the usage of k-means to specifically select for the gold standard differentially expressed genes. Thus, the objective of this work is centered on depicting the most relevant processes altered along viral infection, although more genes might be implicated by their statistical behavior only. Selected genes were used to look at functional biological networks with String tool, including GO-terms and Kegg pathways enrichment analysis, at protein-protein interactions in IntAct database, and to bio interactions by text-mining using PESCADOR. By analyzing 19704 abstracts selected with the PubMed query H1N1 and a list of gold standard genes as concepts in PESCADOR, genes CD47, CLK1, GAPDH and IL8 depicted a scenario comprising 42 connections to biointeractions that are being investigated. Moreover, the webtoolNiji produced by our group was used to represent the differential fold change in all Kegg pathways. All the resultant scenarios will be accessible at http://biodados.icb.ufmg.br/mrvdb (under construction).

17 Apr
1:00pm
P208

The Incidence of Relapse and Remission in LUSC Patients - a genomic and ancestry analysis

Eric David Rebouças de Souza[1], Raquel Pontes Martins[2], Patrick Terrematte[3] and Beatriz Stransky[4]

[1]UFRN - Science and Technology School; [2]Bioinformatic Postgraduate Program; [3]UFRN - Biomedical Engineering Department; [4]Biomedical Engineering Department, Bioinformatics Multidisciplinary Environment

Lung cancer is the second leading cause of cancer death in the United States. It is classified in non-small cell lung cancer and small cell lung cancer, and the former can be divided into three subtypes according to cell origin: squamous cell carcinoma, large cell carcinoma, and adenocarcinoma. According to SEER data (Surveillance, Epidemiology, and End Results), lung and bronchial cancers represent approximately 13.5% of all new cases and around 25% of all deaths from cancer. Lung cancer affects more men than women, particularly black men, with an incidence of 81 cases of black men against 64 cases of white men, per 100,000 inhabitants. In Brazil, Lung, Bronchus and Trachea cancers are the second more incidents and the first cause of death between men (INCA, 2018). Recent large-scale genomic profile studies showed that African American men have higher incidence and mortality rates from prostate and lung cancers, indicating a significant role of the ancestry genetic background in cancer development. Therefore, this study aims to investigate the genetic and clinical alterations that could be related to the different outcomes of relapse and remission events in African-American (AA) and American-European (AE) patients with squamous cell lung carcinoma (LUSC). Clinical and genomic data from 504 patients from the Cancer Genome Atlas (TCGA-LUSC) project were available from the Genomic Data Commons portal (portal.gdc.cancer.gov). Ancestry analysis and classification of these same patients were obtained in the TCGAA (52.25.87.215/TCGAA). Exploratory analysis and raw data preprocessing were done with R scripts. The evaluation of remission and recurrence was done using the Survminer package from CRAN. The analysis of gene expression was done using the R-Peridot tool (www.bioinformatics-brazil.org/r-peridot) and the analysis of functional enrichment was performed with the clusterProfile package from Bioconductor. To investigate the relation of ancestry background and the development and severity in cancer, we plot the cumulative incidence of remission and recurrence events for EA (European American) and AA (African American) subgroups of LUSC patients. The curves showed an inverse behavior of the DFS variable (disease-free status) among the analyzed groups. EA individuals have a 60% probability of becoming disease-free, while AA individuals have a 60% chance of recurrence or progression after 150 months. Based on this finding, differential expression analysis (DEA) was performed between groups with or without remission of the same ethnicity and between groups with the same type of event (remission OR recurrence) of different ethnic groups. The initial analysis identifies 6 underexpressed genes (VSIG2, OR2G6, TMOD3, C3orf62, and STAB1/CL

EVER-1) and 2 overexpressed genes (ELF3 and SDAD1), some of them clearly involved in cell control. These analyses show that the investigation of genetic alterations linked to ancestry in cancer represents an essential step to a better understanding of the cancer biology and can lead to the development of more effective diagnostics and therapies.

17 Apr
1:00pm
P209

Where can we really see the Effect of Neighbouring Nucleotides in Mammals

Fernanda Stussi[1], Lissur Orsine[1], Carlos Xavier[1], Tetsu Sakamoto[2] and Miguel Ortega[1]

[1]UFMG; [2]UFRN

Nucleotide substitution is one of the fuels in the production of genetic variation and throughout evolution. These mutations are supposed to occur randomly along the genome through the time. However, some publications pointed to an effect of the neighborhood pattern of bases on the probability of the occurrence of SNPs. Aiming to investigate comprehensively this event, we built an online database to show the pattern of bases in SNP neighborhood, SNP LANE, after SNP LocAl Neighborhood. In order to obtain the observed frequency we downloaded SNP datasets from five different species of mammals (Mus musculus, Homo Sapiens, Bos taurus, Rattus Norvegicus and Sus scrofa); the FASTA and ASN files from the dbSNP database were downloaded and parsed with in-house Perl scripts according to the genomic region the SNP belongs to (intron, CDS, 5UTR, 3UTR and intergenic regions). The data has been filtered for the purpose of removing indels and those SNPs ambiguous for more than two bases. In the same database for each entry, the sequence information present in the FASTA file was concatenated with the region information that is present in the ASN file for determination of baseline frequencies flanking pair of bases, such as A/T, to compare with the analogous SNP. For each SNP class nucleotide frequencies were calculated for the first five positions upstream and downstream of the SNP position in the genome. The expected baseline nucleotide frequencies for the positions neighboring the SNP were estimated by randomly choosing positions in the genome flanking pairs of bases and retrieving the nucleotides surrounding it. For the purpose of contrasting the observed frequency with the baseline frequency we applied the chi-square test for each position. From 901,829,443 SNPs downloaded, 392,591,139 (44%) passed through our quality filters. They were then classified by their organism, type of substitution (e.g. W or A/T transversion), genomic region and position surrounding the mutation site, adding up to 1200 combinations. Remarkably, in the majority of the cases the baseline frequency was not significantly different from the observed data, indicating that the genome is not randomly composed and its composition generates most of the observed and elsewhere reported neighboring bias. In just 92 of those categories (7.6%) the difference between the observed frequency in mutation sites and baseline frequency were significantly distinct according to the chi-square test. The majority of it, 56 cases (4,6) were found in Bos taurus, an organism which in a previous analysis of the proportions of transitions and transversions was noticed a bias that changed the relative proportions between these categories, suggesting that sampling might not be random in the cow project. With the exception of cow data, in only 3% of the cases the baseline frequency

does not explain the neighboring bias. Analysis of the graphics for the neighboring patters show remarkable similarity between observed and baseline frequencies. In order to investigate CpG effect on the origin of the bias, we first deaminated all remaining C in CpG and observed a small increase in the bias. Not only that, different percentages of amination of "CpA" and "TpG" back to CpG dinucleotides along the genomic regions were simulated. We have determined that the bias is completely erased by certain percentages of amination. It was noteworthy that between 25% to 35% of amination K, M, W, S substitutions showed an Euclidian distance from the fifths positions to the firsts positions lower than 0.02 - this means that the difference in frequency distribution of the bases from the most distant to the nearest became very small, therefore we were unable to see the neighboring nucleotide effect on those conditions, suggesting that this bias is observed mostly due to deamination of C in CpG. On the other hand, R and Y substitutions did not respond to amination. In conclusion, it is suggested that dinucleotide composition produces the previously reported neighborhood bias on SNP probability. Most of this effect might be explained by the deamination of C in CpG and we suggest that originally genome would have 25 up to 35% of the present CpA and TpG in the form of CpG.SUPPORT: CAPES BIOLOGIA COMPUTACIONAL - BSC NETWORK and FAPEMIG

17 Apr
1:00pm
P210

Bioinformatics applied to industry as a tool for delineating pharmaceutical products with the mixture α, β-amirenone and cyclodextrins

Gabriela Suassuna Bezerra[1], Luana Carvalho Oliveira[1], Thalita Sévia Soares de Almeida[1], Elayne Barros Ferreira[1], Anna Thereza Sousa Queiroz[1], Paulo Henrique Santana[1], Euzébio Guimarães Barbosa[1], Emerson Silva Lima[2], Valdir Florêncio Veiga Júnior[3] and Ádley Antonini Neves de Lima[1]

[1]UFRN; [2]UFAM; [3]Instituto Militar de Engenharia-RJ

The plants of the family Burseraceae have genera, such as Protium, which are producers of a resin that presents oily, viscous and amorphous characteristics and consists mainly of pentacyclic triterpenes, among which are α-amirenone and β-amirenone, that as the other triterpenes are generating a lot of interest, given their pharmacological profiles. The therapeutic activities that can be cited are anti-inflammatory, analgesic, healing, expectorant, antiulcerogenic, antifungal, antimicrobial and antioxidant. They also present hypoglycemic and hypolipidemic actions, thus having positive effects on metabolic disorders. However, the physicochemical characterization of blend demonstrates a crystalline nature in view that showed a well-defined fusion event and intense crystalline reflections, probably exhibiting a low aqueous solubility, although promising for the potential development of new drugs, it is necessary, alternatives to bypass this obstacle. Among all the alternatives, the performance of the cyclodextrins (CD), which are natural cyclic oligosaccharides (with six, seven, eight or more α 1,4-glucopyranose units) were evaluated, presenting a truncated cone shape with a hydrophobic cavity and external surface hydrophilic due to the presence of hydroxy groups. These, when substituted, give rise to the derivative CDs and demonstrate improvements in solubility. Such characteristics make them of great interest for the development of pharmaceuticals from complexes with poorly water-soluble drugs. The use of the cyclodextrins as excipient becomes viable, so that they act increasing the solubility and dissolution of the poorly soluble drugs, stability, safety and their bioavailability. This may lead to a decrease in the dose administered for medicinal products having the same possibility of biological activity. As well as, the cyclodextrins can provide a better drug release from the pharmaceutical forms, which makes them release promoters. In order to guarantee the formation of the complexes it is necessary to establish the stoichiometric ratio between the drug and the CD used, in such a way as to achieve a better complexing efficiency and dissolution rate. However, methods for predicting the stoichiometric ratio have traditionally been used, which are costly and require a lot of bench time. Molecular docking evaluates intermolecular interactions by means of sufficiently precise algorithms, in addition to simulating the binding energy between molecules, becoming important in the development of drugs against the reduction of costs with traditional experiments, avoiding flawed investments and shorter research

time, generating better cost-effectiveness for pharmaceutical companies in the initial stages of the design of new products. Therefore, bioinformatics becomes increasingly relevant for the development of new drugs. In this context, the objective of this work was to predict, by means of in silico study, the formation of inclusion complexes between the triterpenic mixture α, β-amirenone with natural and derived cyclodextrins, besides evaluating the best complex formed in relation to free energy of the bond obtained during molecular docking. The 3D models of α-amirenone and β-amirenone ligands were drawed using MarvinSketch software 18.16. These models were submitted to docking simulations with four cyclodextrins: hydroxypropyl-γ-CD (HPγCD), hydroxypropyl-β-CD (HPβCD), γ-CD and β-CD. All complexes were optimized by PM6-DH+ semi-empirical theoretical method using MOPAC software. The results of the conformations these complexes (cyclodextrins-binder) were analyzed visually in the UCSF-Chimera software. The complexes binding energy with the triterpenic mixture was defined as energies mean obtained for more stable conformations of models separately. The complexes with natural CDs (β and γ) were simulated in stoichiometric ratios (CD: Drug) 1:1 and 2:1. The binding energies obtained for β-CD (1:1), β-CD (2:1), γ-CD (1:1) and γ-CD (2:1) complexes were -40.24 kcal/mol-1, -109.67 kcal/mol-1, -21.07 kcal/mol-1 and -81,875 kcal/mol-1, respectively. The complexes with the derivative CDs (HPβCD and HPγCD) were simulated only in the ratio 1:1, since in the ratio 2:1 complexation was not favorable. The binding energies obtained for the HPβCD and HPγCD complexes were -26,095 kcal/mol-1 and -21,275 kcal/mol-1. Considering the results, the complex formed in the 2:1 ratio with β-CD showed the lowest binding energy after the docking simulations, suggesting that it would be the best inclusion complex to be used in the development of a new product. Thus, bioinformatics can be used with ally in the research of new pharmaceuticals products, considering that, starting from docking simulations, it was possible to predict the formation of inclusion complexes with different cyclodextrins (natural and derived) and the triterpenic mixture α, β-amirenone, as well as to evaluate the more favorable stoichiometric ratio.

17 Apr
1:00pm
P211

The effect of extended phenotypes in the overall fitness of populations: testing the Extended Fitness Hypothesis

Guilherme Fernandes de Araujo and Sandro Jose de Souza
UFRN

Graph-based biological simulations of populations have been used to study the fixation of genes that offer reproductive advantages to competing individuals in a same environment. The selection coefficient, higher in individuals with advantageous characteristics, leads to a faster and higher chance of fixation of the gene that promotes it. Previous works have shown that extended phenotypes can be modelled as attached characteristics of individuals, providing reproductive bonuses and increasing the chance to fix their extended phenotype in a population. These phenotypes can be reused by other capable individuals after the death of the previous user, since it is a constructed environmental feature and not an individual characteristic. The effect of extended phenotypes in the fitness of a given population has been theoretically explored by the "Extended Fitness Hypothesis". A new software written in C++ has been developed to simulate a population in which extended phenotypes are available to conspecifics and their effect in the population fitness is evaluated. Simulations were run on a Barabasi-Albert graph with 500 nodes, 50/50 rate for producing and non-producing individuals, where half of the producing individuals starts the simulation with an extended phenotype attached. A limit of 5000 cycles was set, and a phenotype usage bonus of 0.15 was defined. Experiments ran on expiry time settings of 3, 5, 10, 20, 30, 40, 50 and 100 cycles, and a 4% population renewal rate, where 4% of individuals die and are born in any cycle. Comparison between population containing or not extended phenotype producers allowed us to find that extended phenotypes have a strong effect on the fitness of given population. If confirmed, such results would give a strong support to the "Extended Fitness Hypothesis".

Partial sequencing of plasmids from Chromobacterium violaceum Brazilian strains

17 Apr
1:00pm
P212

Daniel C. Lima[1], Inácio Gomes Medeiros[2], Rita de Cássio Silva-Portela[3], Francisco Carlos da Silva Junior[3], Jorge Estefano Santana de Souza[4], Lucymara Fassarella Agnez-Lima[3] and Silvia R. Batistuzzo de Medeiros[3]

[1]Instituto Federal de Educação, Ciência e Tecnologia do Rio Grande do Norte, Campus Ipanguaçu, Brazil; [2]Programa de Pós-Graduação em Bioinformática, Instituto Metrópole Digital, Universidade Federal do Rio Grande do Norte, Natal, Brazil; [3]Laboratório de Biologia Molecular e Genômica, Centro de Biociências, Universidade Federal do Rio Grande do Norte, Natal, Brazil ; [4]Bioinformatics Multidisciplinary Environment, Instituto Metrópole Digital, Universidade Federal do Rio Grande do Norte, Natal, Brazil

Chromobacterium violaceum (a.k.a. C. Violaceum) is a Gram-negative ß-proteobacteria with important biotechnological properties and is associated to hard-to-treat infections in immune-depressed individuals. The Malaysian strain ATCC 12472 was sequenced in 2003 and more recently, Lima and colleagues have discovered a plasmid in this strain comprising of phage, plasmid partitioning and hypothetical proteins genes. In the same work, they have demonstrated from restriction pattern analysis that this plasmid could be present in other strains isolated from Brazil. We report in this work the sequencing of six plasmids isolated from these strains, demonstrating that pChV1 first discovered at C. Violaceum ATCC 12472 is also present in Brazilian strains. Plasmids strains were sequenced by NGS using Ion Torrent PGMTM System (Life Technologies). Sequenced libraries were assembled with SPAdes v3.10.1. Redundans v. 10/2014 was used for removing gap regions and CAP3 software, v. 08/2013, for final refinement. ORFs that are annotated in NCBI Nucleotide's C. Violaceum genome, entry NC_005085.1, were mapped to assembled scaffolds via BLAST, version 2.6.0+, with maximum e-value of 1e-5. A total of six assemblies of Brazilian strains plasmids were obtained, with the following genome coverage, respectively: 67%, 50%, 41%, 30%, 4%, and 2%. All assemblies had about 260,000 reads with varying number of them aligned with pChV1. The number of pChV1 ORFs found in each assembly were respectively: 27, 21, 10, 7, 2 and 0. First two strains, named CVACO5 and CVACO2, are believed to probably to have pChV1, due to the number of reads aligned against it, horizontal genome coverage, number of ORFs from pChV1 as well as from the restriction profile of the plasmids described at previous work. Regarding the ORFs found on each strain, we observed that most strains had ORFs coding for plasmid replications proteins as ParA and RepA. Moreover, strains CVACO5 and CVACO2 presented most ORFs from pChV1, as expected. These results shows that pChV1 plasmid is present not only in C. Violaceum ATCC 12472, but also strains isolated in the Amazon Region in Brazil. Thus, there is still a huge biotechnological potential to be explored for C. Violaceum, because of the presence of this new plasmid.

IN SILICO CHARACTERIZATION OF PROTEINS WITH UNKNOWN FUNCTION FROM CHROMOBACTERIUM VIOLACEUM

17 Apr
1:00pm
P213

Jonathas Diego Lima Santos, Augusto Monteiro de Souza and Silvia Regina Batistuzzo de Medeiros

Universidade Federal do Rio Grande do Norte

Chromobacterium violaceum is a free-living Gram-negative β-proteobacteria, which has been found in water and soil in tropical and subtropical regions. This bacterium has high biotechnological value due to its inherent antitumoral, antifungal, antibiotic, antiviral and antileishimania activities. It also possesses bioremediation properties of contaminated sites by heavy metals, production of bioplastics and adaptation to environments with high incidence of UV radiation. After having its genome sequenced, it was verified that most of its ORFs still do not have their elucidated function, being necessary studies that aim at this objective. Thus, the objective of this study was to characterize proteins with unknown function of C. violaceum grown under simulated microgravity conditions. Proteins were obtained by the shotgun proteomics technique after cultivation of C. violaceum in a Rotator Culture Cell System (RCCS4), a reduced gravity simulator, grown for 5 and 12 hours at a speed of 40 rpm. After proteomics analysis, 30 proteins with unknown functions were identified and 20 proteins were selected for in silico characterization. First, homology searches were performed to identify groups of organisms that have the protein of interest and to report whether it is conserved in different genera. Through online database DOOR (http://csbl.bmb.uga.edu/DOOR/index.php), the genomic context of these proteins was analyzed in order to verify if their neighboring ORFs had a known function and be able to infer a possible function. Pfam (http://Pfam.xfam.org) allowed to verify if these proteins are conserved hypothetical ones. In addition, protein-protein interaction networks were built using the online tool STRING (version 11.0) and Cytoscape software (version 3.4.1) to identify which known proteins of C. violaceum interact with the proteins of interest, and thus identify the function and the metabolic pathways where these proteins could be involved. After careful analysis, all proteins from the 20 analysed showed to be conserved hypothetical, being, probably CV_2933 a peptidyl-prolyl isomerase; CV_4364 a Copper chaperone PCu (A) C; CV_1173 a cytochrome-c oxidase; CV_1539 and CV_4329 proteins of the ABC transport system that recognize the broad range of ligands from metal ions to amino acids, sugars and peptides; CV_2258 and CV_3947 enoyl reductases acting on fatty acid biosynthesis; CV_3008 a signal receptor protein in regulatory response processes; CV_2019 a aldehyde dehydrogenase; CV_3961 a methyl phosphate synthase-like that acts on the synthesis of methane; CV_1546 a 3-oxoacyl- [acyl-carrier protein] reductase acting on fatty acid metabolism; CV_0018 a biotin carboxylase (a component of acetyl CoA carboxylase)

that also participates in fatty acid biosynthesis; CV_0175 a cytochrome C oxidase; CV_1182 a endoribonuclease L-PSP enzyme-like that acts on the biosynthesis of aromatic amino acids; CV_3910 a holliday junction resolvase with an important role in the segregation of the chromosomes; CV_0971 an enzyme of Fumarylacetoacetate (FAA) hydrolase Family; CV_3869 a bacterial DNA-binding protein important in the transcription process; CV_3374 and CV_1366 from the phasin family protein, which are involved in PHA granule stabilization and distribution to daughter cells upon cell division; CV_3977 and CV_1891 OmpA family proteins with roles in bacterial pathogenesis and immunity; CV_0868 protecting protein against oxidative stress by destroying radicals that are produced inside cells and which are toxic to biological systems. This preliminary prediction is extremely important as it is a pre-requisite for elucidating roles that these proteins play in biological systems. It is necessary to increase these results with data of tridimensional structures and molecular dynamics to provide subsidies for in vitro functional characterization of these proteins by complementary assays and specific activity of these proteins.

MODEL OF MIXTURE OF DISTRIBUTIONS FOR THE EFFECTS OF GENETIC CHANGES IN THE TUMOR PROGRESSION OF BREAST CANCER

17 Apr
1:00pm
P214

Josivan Ribeiro Justino, André Luís Fonseca Faustino, Clóvis Ferreira dos Reis, Danilo Lopes Martins, Beatriz Stransky, Jorge Estefano Santana de Souza and Sandro José de Souza

UFRN

The identification of genes with bimodal expression patterns may contribute to the understanding of important biological processes. Among the possible hypotheses of investigation, it is believed that this bimodal pattern can reveal molecular signatures that distinguish the tumor subtypes, which would contribute to a better understanding of the evolutionary dynamics of cancer, besides being part of the search for critical clinical targets for diagnosis and treatment of tumors (MOODYA et al., 2019). Thus, this study aims to verify if the bimodal expression of a given gene in breast cancer patients correlate with the underlying biological processes associated with a better or worse prognosis of the disease. Gene expression data from 1058 patients with primary breast tumor were obtained from cBioPortal for Cancer Genomics, in a total of 20,305 genes. Twenty-one genes with bimodal expression were identified from the probabilistic Gaussian Mixture Models (GMM) model (STYLIANOU et al., 2005), and the patients identified at each of the two levels of expression were sampled. The bimodal were characterized into according to their recognized functions. The patients identified according to the levels of the bimodal genes were classified according to the presence or absence of Estrogen Receptor (ER) and Progesterone Receptor (PR) and HER2 in 4 breast cancer subtypes: Luminal A, Luminal B, HER2 positive or Negative Triple (for estrogen, progesterone, and HER2 receptors). Survival analyses showed a significant difference of the patients identified with the Negative Triple subtype of breast cancer of group 2 (high expression) to group 1 (low expression), for 15 of the 21 bimodal genes. In the next steps, we will perform Differential Expression Analysis (DEA) for the groups identified in survival analysis to better investigate the influence of the bimodal genes expression into the biological processes, pointing out the possible associations of expression levels (high and low) with the regulation pathways of the genes that presented bimodal distribution.

On the impact of the pangenome and annotation discrepancies while building protein sequence databases for bacteria proteogenomics

17 Apr
1:00pm
P215

KARLA CRISTINA TABOSA MACHADO[1], Suereta Fortuin[2], Gisele Guicardi Tomazella[1], André Faustino Fonseca[1], Rob Warren[2], Harald Wiker[3], Sandro José de Souza[1] and Gustavo Antônio de Souza[1]

[1]Bioinformatics Multidisciplinary Environment, Universidade Federal do Rio Grande do Norte, Brazil; [2]DST/NRF Centre of Excellence for Biomedical Tuberculosis Research/SAMRC Centre for Tuberculosis Research, Division of Molecular Biology and Human Genetics, Faculty of Medicine and Health Sciences, Stellenbosch University, South Africa.; [3]The Gade Research Group for Infection and Immunity, Department of Clinical Science, University of Bergen, Norway

In proteomics, peptide information within mass spectrometry data from a specific organism sample is routinely challenged against a protein sequence database that best represent such organism. However, if the species/strain in the sample is unknown or poorly genetically characterized, it becomes challenging to determine a database which can represent such sample. Building customized protein sequence databases merging multiple strains for a given species has become a strategy to overcome such restrictions. However, as more genetic information is publicly available and interesting genetic features such as the existence of pan- and core genes within a species are revealed, we questioned how efficient such merging strategies are to report relevant information. To test this assumption, we constructed databases containing conserved and unique sequences for ten different species. Features that are relevant for probabilistic-based protein identification by proteomics were then monitored. As expected, increase in database complexity correlates with pangenomic complexity. However, Mycobacterium tuberculosis and Bortedella pertusis generated very complex databases even having low pangenomic complexity or no pangenome at all. This suggests that discrepancies in gene annotation is higher than average between strains of those species. We further tested database performance by using mass spectrometry data from eight clinical strains from Mycobacterium tuberculosis, and from two published datasets from Staphylococcus aureus. We show that by using an approach where database size is controlled by removing repeated identical tryptic sequences across strains/species, computational time can be reduced drastically as database complexity increases.

17 Apr
1:00pm
P216

Systems biology approaches in the investigation of bottlenecks in KEGG pathways

Leonardo René dos Santos Campos, Diego Morais, Igor Brandão, João Vitor Cavalcante, Jorge Estefano Santana de Souza and Rodrigo Juliani Siqueira Dalmolin
UFRN

The ever-increasing availability of sequencing data poses several challenges to researchers aiming to analyze functional profiles of sampled organisms as the process of reliably mapping the sequences to databases, extracting functional annotations, and visualizing the information obtained is a complex and delayed one. Traditional sequence alignment tools like BLAST could spend several days of computing time to return matches against large databases (e.g. NR or TrEMBL) for typical metagenomic datasets queries consisting of thousands of sequences. The employment of fast aligners like DIAMOND and PALADIN is essential for rapid functional analysis of today's data. Notably, the eggNOG-Mapper tool, which uses DIAMOND in the sequence mapping step, provides accurate functional assignments for Gene Ontology terms, COG functional categories, and KEGG pathways. KEGG database is a great source of curated functional information. It holds a knowledge base on pathways maps of molecular interaction and orthology relationships between genes/gene products, providing high-quality reference pathways maps regarding thousands of organisms with complete genomes. Each KEGG pathway has a corresponding XML file describing it that can be represented as a graph where nodes are enzymes and links represents reactions between enzymes and metabolites. That graph structure enables the use of systems biology approaches aiming to infer essential nodes in the networks, which, combined with mapped sequences from omics samples could lead to the identification of functional signatures. In this context, we developed a procedure to obtain and process KEGG pathway data using R packages from Bioconductor's and CRAN's repositories. We used KEGGREST package to load the reference pathway of interest from the REST API provided by KEGG. The KGML (KEGG XML) file is processed and transformed into a graph object whose nodes are KEGG EC numbers. We used the iGraph package to extract several features from the graph like communities, betweenness, closeness, and clustering. Our main interest is the bottlenecks, obtained through the articulation points of the graph. Bottlenecks are key nodes that play an important role in metabolic processes described by the pathways. We have selected the Glycolysis/Gluconeogenesis reference pathway as a simple example for applying the procedure and we have identified two nodes as bottlenecks of that pathway, matching the visual evaluation of the same pathway in KEGG. Concluding the procedure, we provide a visualization of the pathway marking the bottlenecks selected using pathview R package. Further analysis for curating the importance of identified bottlenecks are necessary. We are selecting key pathways with well-described metabolism problems to assess the robustness of our

procedure, like the valine, leucine and isoleucine degradation reference pathway, and the tyrosine metabolism reference pathway.

IN SILICO EVALUATION OF TYROSINE KINASES INHIBITORS WITH PHARMACOLOGICAL POTENTIAL IN ERYTHROLEUKEMIC CELL LINES

17 Apr
1:00pm
P217

Lucas de Sousa Pontes, Caroline Aquino Moreira-Nunes, Júlio Paulino Daniel, Felipe Pantoja Mesquita , Maria Elisabete Amaral De Moraes, Manoel Odorico De Moraes-Filho and Raquel Carvalho Montenegro

Pharmacogenetics Laboratory, Drug Research and Development Center (NPDM), Department of Physiology and Pharmacology, Federal University of Ceara, Fortaleza, CE, Brazil.

Introduction and objectives: The remission of Chronic Myelogenous Leukemia (CML) by tyrosine kinase (TK) imatinib mesylate, a BCR-ABL inhibitor, is highly effective, however, in resistant cases this treatment may not be enough, due to the participation of others TKs pathways. Responsible for the major intracellular signaling mechanism, the TKs act promoting cellular functions such as growth, proliferation, motility, among others, in different kinds of cancer, becoming the most focused pathway in the development of inhibitory drugs. In silico predictions can be used for the accomplishment of clinical studies looking for potential therapeutic effect for several molecules, searching for targets and off-target through a wide database of already described molecular interactions. The aim of this study was to identify new targets for biologically tested tyrosine kinase inhibitors (TKIs) with therapeutic efficacy in the erythroleukemic cell line K-562 comparing the data obtained by the prediction algorithms with those obtained experimentally in vitro. Material and methods: A set of 367 molecules present in a database of small ligands were evaluated, and from that set, 30 compounds was filtered based on the lowest GI50 concentration and cytostatic activity in the erythroleukemic cell line K-562. The database search tools PubChem and ChEMBL was used to identify information's such as: compound identification, molecular structure, biological interactions and SMILES (Simplified Molecular Input-Line Entry System). The canonical SMILES linear notation obtained was used in the Swiss Target Prediction online tool for target analysis based on structural similarity from molecules with experimental data. The Protein Data Bank (PDB) were used to obtain the structural model from predicted targets. All data analysis and visualization were performed using R programming language and integrated development environment for statistics. The molecular docking using Vina was performed for 30 candidates against five TKs. To check for functional association of the protein-protein interactions (PPI) of the targets were mapped using STRING database, only interactions with a combined score > 0.7 were considered significant for this study. Results: A total of 15 targets were obtained for each compound, however, only those pointed out as probability of interaction score greater than 0.8 were selected, reaching a total of 220 interactions. Subsequently, the molecular interactions already confirmed in online databases and by biological experimentation by Dr. Jonathan Elkins et al. were discarded, reaching a total of 72 predicted interactions. The PDB IDs for

the predicted targets CDK1, CDK6, IMPDH2, MPK8 and PIK3CD were as follows: 4YC3, 1XO2, 1NF7, 1UKI and 5DXU. As preliminary result, compounds GW809897X, GW589933X, GW810372X, GW778894X and GSK1326255A seems to be the best for each one of the targets. The analysis of PPI networks revealed that a lot of these predicted targets are TK and could interact with each other. Compounds GW809897X and GW778894X inhibits FLT1, FLT4 and KDR cell surface proteins that plays an essential role in angiogenesis regulation. Furthermore, Compounds GW589933X, GW810372X and GW778894X showed potential to inhibit several Cyclin-dependent kinases (CDK) which have an important role regulating the cell cycle. Data showed that these three compounds has potential and the ability of also inhibit GSK3A and GSK3B, TKs that act as a negative hormonal regulator of glucose homeostasis. Was predicted for compound GSK1326255A potential to inhibit several carcinogenic promoters' proteins of MAP kinase subfamily. Conclusion: Five compounds demonstrated potential in the inhibition of important proteins in carcinogenic processes, moreover, GSK1326255A evidenced as a promising inhibitor for hematologic malignancies since MAP kinase subfamily plays an important role mediating related pathways. MAP kinase subfamily inhibition seems promising, since these TKs have a close relationship with CML, with increased activity after BCR-ABL inhibition by imatinib mesylate. Several steps have yet to be taken to obtain a conclusive result on the predicted molecular interactions, starting with in vitro experiments.

17 Apr
1:00pm
P218

Role of G-quadruplex forming sequences on methylation inside CpG islands

Manuel Jara-Espejo and Sergio RP Line
Faculdade de Odontologia de Piracicaba - UNICAMP

Introduction

The normal development of an organism requires precise interactions between genetic and epigenetic mechanisms. DNA methylation plays crucial roles in mammalian development, silencing genes and contributing to embryonic stem cell differentiation and cell fate determination. Since G-quadruplexes forming sequences (G4FS) tend to be enriched at DNA regulatory regions, like promoters, they may participate on definition of CpG dinucleotide methylation pattern inside CpG islands.

Objective To investigate the role of G4FS on CpG methylation patterns of CGI.

Material and Methods Using the Roadpmap Epigenomics Project database, we obtained Whole genome bisulfite sequencing (WGBS) methylation data at single-nucleotide resolution. The data corresponded to one Embryonic stem cell (ESC.HUES64) and three derived embryonic germ layers. We selected methylation data only for CpGs mapping inside a CGI (CGI/CpG). G4FS features were obtained using an in-house algorithm. Using the RNAfold tool we calculated the Minimum free energy (MFE) for each G4FS identified; Mfe value was used as a measurement of G4FS stability. Then, G4FS were divide into two groups based on its Mfe value: low and high stable G4. Distance between each CpG and its nearest G4FS were calculated. Additionally, we downloaded DNase I hypersensitivity sites (DHS) peak coordinates and we identified G4FSs and CpGs overlapping DHS sites. We used a sliding window script along the distance between CpGs and G4FS, and simultaneously Wilcoxon Signed-Rank tests were performed between the CpG methylation values when CpGs were grouped based on nearest G4 stability (Mfe).

Results We found that CGI/CpGs exhibited low methylation when located inside CGIs having a G4FS; moreover, the lowest methylation values were associated with G4FS located within CGI and mapping open chromatin regions. So, the presence of higher folding potential G4FS within the CGI seems to be related to a low or unmethylated pattern. CGI/CpGs mapping open chromatin exhibited a broad low methylated state. Interestingly, high stability G4FS were associated with lower methylation when compared to low stability G4FS. Moreover, high stability G4FS were more effective in promoting hypomethylation when they were within CGI. CGI/CpGs located in closed chromatin regions tend to be more methylated regardless of its location inside or outside CGI. Besides, CGI/CpGs associated with high stability G4FS were less methylated, and this effect was even more evident when the G4FS were located outside CGI. The effect of the distance between each CGI/CpG and its nearest G4FS was also analyzed. Using a sliding window approach, we found that CGI/CpGs located close to high stability G4FS were less methylated than those located close to low stability G4FS; as the

distance increased, the methylation remained low or increased only at higher distances for high stability G4FS. Therefore, the protective effect of G4FS against methylation is not only dependent on its stability and chromatin accessibility but also on its distance to CGI/CpG and its location (inside or outside CGIs).

Funding Agency: CNPQ grant Nº 141004/2018-5

17 Apr
1:00pm
P219

Systems Analysis of Rhesus Monkeys Acutely Infected with Zika Virus

Mariana Araújo-Pereira[1], Vinicius Maracaja-Coutinho[2] and Helder I. Nakaya[1]
[1]University of São Paulo; [2]Universidad de Chile

Zika Virus (ZIKV) is an arbovirus that has gained high relevance in recent years. Although ZIKV infection generally induces mild symptoms, in some cases, the infection is associated to congenital syndrome, including microcephaly in newborns and Guillain-Barré syndrome in adults. Long non-coding RNAs (lncRNAs) play a key role in modulating the immune system but nothing is known about their function in acute ZIKV infection. Non-human primate models offer a great opportunity to study acute infection with ZIKV. We have infected 4 Rhesus monkeys with a Brazilian ZIKV strain and collected their blood before and after 1, 3, 5, 7, 10 and 14 days of infection. We then performed RNA-seq to assess the transcriptional changes in the blood during acute infection. Our systems biology analyses revealed a regulation of immune response during the infection including the identification of many lncRNAs that were induced upon ZIKV infection. We identified around 9.210 lncRNAs, 3.246 of which weren't annotated on the ENSEMBL database, and 140 of which were differentially expressed in some comparison. Some genes associated with immune response such as STAT1, JAK1, IFNGR2, SMG6 and DCP1A, appear to be negatively regulated. Co-expression modular analyses and network analyses have revealed the putative pathways and genes affected by these lncRNAs. Such as, TCONS_00122811 and TCONS_00021967, lncRNAs not yet deposited in public banks, were highly correlated with IL12B and IGFBP2, respectively. These lncRNAs can potentially regulate the expression of the genes mentioned previously, a hypothesis that needs more research. With this study we found many previously undescribed lncRNAs in the Macaca mulatta genome, some of which are correlated with important genes for immune response against viral infection. We have also compared these results with several human RNA-seq studies of neuronal cells infected with ZIKV and found that some human lncRNAs may play similar roles to those found in our non-human primate study. Overall, our results suggest that ZIKV modifies the expression of coding genes and lncRNAs as a mechanism of cellular evasion.

Systems immunology to predict regulation factors on skin cutaneous melanoma

17 Apr
1:00pm
P220

Mindy Stephania Muñoz Miranda and Helder Takashi Imoto Nakaya
Computational Systems Biology Laboratory, Universidade de São Paulo (CSBL-USP)

Introduction and Objectives: Cutaneous melanoma is a melanocyte skin cancer and 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 Gene Expression Omnibus (GEO) is a repository of high throughput gene expression data which contains thousands of melanoma samples. Previous studies performed Bulk RNA-Seq analysis with skin cutaneous melanoma data from The Cancer Genome Atlas (SKCM-TCGA). Despite this, there are currently no publicly available melanoma single cell co-expression studies. On the other hand, few studies have applied systems biology to investigate the molecular mechanisms of the immune system against melanoma, melanoma evasion from the immune system and its proliferation in other tissues. Therefore, the aim of this study was to perform a single cell transcriptomic analysis to understand the gene co-expression of blood cells from metastatic melanoma patients. Material and Methods: Normalized single cell gene expression data was obtained and separated by cell types. For each cell type CEMiTool was applied to find co-expression modules and their respective enriched pathways. Module genes from immune-related pathways were used in the X2K-Web tool to infer upstream regulatory networks, thus combining co-expression networks, transcription factor enrichment analysis, protein-protein interaction with kinase enrichment analysis. The analysis for each type of immune-cell was performed to find related genes to melanoma. Results: Enrichment in negative regulation of cell proliferation pathway was observed in one of the macrophage co-expression modules, of which some genes are common transcription factors previously related to metastatic tumor formation and drug resistance. The CDKN1A gene, usually up-regulated in malignant melanoma cells, was identified as a hub gene in Natural Killer cell co-expression networks. Conclusions: Our analysis described transcription factors in immune cells in melanoma patients, some already related to drug resistance in the melanoma progression treatment. These results show the link between the immune system and melanoma progression which could explain drug resistance in immunotherapies.

17 Apr
1:00pm
P221

Application of The Noisy Neighbors Algorithm in a Multiomics Pipeline to study Multiple Sclerosis

Rafael Mendonça Colares[1], Suzana Pôrto Almeida[1], Francisco Eder de Moura Lopes[1], Raul Victor Medeiros da Nóbrega[2], Saul Gaudencio Neto[1], Leonardo Tondello Martins[1], Ana Cristina de Oliveira Monteiro Moreira[1], Antonio Carlos da Silva Barros[2], Pedro Pedrosa Rebouças[2], Victor Hugo Costa de Albuquerque[3] and Kaio César Simiano Tavares[1]

[1]Laboratório de Biologia Molecular e do Desenvolvimento, Universidade de Fortaleza, Fortaleza, Ceará, Brazil; [2]Laboratório de Processamento de Imagens e Simulação Computacional, Instituto Federal de Ciência e Tecnologia do Ceará, Fortaleza, Ceará, Brazil; [3]Programa de Pós-Graduação em Informática Aplicada, Universidade de Fortaleza, Ceará, Brazil

INTRODUCTION AND OBJECTIVES Genome-wide Association Studies (GWAS) have identified innumerous Single Nucleotide Polymorphisms (SNP's) associated with human complex traits like Multiple Sclerosis (MS). Most of the disease-associated genetic variants are noncoding, becoming challenging to identify their consequences. However, with the advance of epigenomics, it is possible to use a combination of mapping technologies to find, among non-coding SNPs, the ones inside regulatory regions. With these regulatory SNPs, it is possible to use known Transcription Factor (TF) binding profiles databases to find which TFs are binding in the regulatory SNPs region, therefore inferring their biological purpose. Knowing that SNP's can be inherited in haplotypes, a haplotype-based approach used by Khankhanian et al.(2015) identified 932 relevant SNP-haplotypes of previously single-SNP GWAS for Multiple Sclerosis (MS). Thus, the aim of this work was to investigate the 932 SNP-haplotypes for MS in a pipeline associated with the algorithm we developed (Noisy Neighbours Algorithm – NNA) to identify which of them are inside regulatory regions, and use these regulatory SNP-haplotypes to find known TFs that binds differently depending on the allele, therefore showing which transcription factors are related to MS and validating Noisy Neighbours Algorithm.

MATERIAL AND METHODS The NNA was used to identify Khankhanian's SNP-Haplotypes genomic location, major and minor alleles for each SNP in dbSNP database and to create blocks for each SNP-haplotype, considering a size of 32 nucleotides from each block as the TF binding site possible length, and all possible combinations of the occurrence of an SNP in a block with the purpose of showing differential TF binding in regulatory region affected by the event of more SNPs per block. After that, the SNPs were analyzed with the Roadmap Epigenomics data and Chromatin States Models to select the Haplotype-SNP that were inside a regulatory region. Then, the JASPAR TF binding profiles database was used to identify possible TFs that would bind in the regulatory SNP-Haplotype region, and returns its findings output in a table.

RESULTS AND DISCUSSION Eight TFs (GATA-1, GATA-2, GATA-3, GATA-4, GATA-5, GATA-6, HLTF, SRF) were found to bind to a sequence block combined with the SNPs rs12751613 and rs12122806, which are inherited in haplotypes and are present inside enhancer region in the following cells types or tissues, according to Roadmap Epigenomics: Brain Anterior Caudate, Brain Cingulate Gyrus, Brain Prefrontal Lateral Cortex, Primary T helper 17 cells PMA-I stimulated, Primary T Cells from Peripheral Blood, Primary mononuclear cells from peripheral blood, Fetal Thymus, Primary Hematopoietic Stem Cells, Primary Natural Killer cells and neutrophils from peripheral blood. It is noteworthy that only the TFs GATA1 and GATA2 would appear as a result without the implementation of the Noisy Neighbors Algorithm. Of these 8 TFs, only GATA-3 and SRF were found in scientific literature associated to MS, and they were related as a protective factor of MS and a neuroprotective function, respectively. Being GATA-3 related to the increase of IL-4 levels and consequently to the increased Th2 cells response, as they have been proposed to play a neuroprotective role in MS, and the neuronal SRF exerting neuroprotective functions including axon growth and pathfinding as well as motor neuron survival and regeneration. According to Khankhanian's data, the MS haplotype-risk SNPs rs12751613 and rs12122806 have the alleles C (minor allele) and T (wild-type), respectively. Thus, the results showed that, when both risk alleles are present, GATA-1 tends to be the only one to bind, which allows us to infer that its activity increases the pathogenesis of MS. On the other hand, when inherited for both SNPs C alleles (Minor allele) and C (Minor allele), all 8 TFs could connect to the regulatory region. Even though there is no information over the others TF's relating to MS, it is known that all members of the GATA transcription factor family bind to relatively equal nucleotide sequences. It's possible that in different instances with alleles other than C and T inherited for the SNPs rs12751613 and rs12122806, there may be competition for the GATA Family and different interaction with SRF and HLTF transcription factors over the genome regulatory region associated.

CONCLUSIONS In conclusion, this study showed the capability of the data integration involving GWAS, epigenomics and regulomics associated with the NNA to identify new pathways to functional consequences in complex diseases. In addition, there is still need to validate the competition and different transcriptional regulatory interactions between TFs related to MS from this study with more experimental research.

Funding Agency: CNPq - Conselho Nacional de Desenvolvimento Científico e Tecnológico CAPES – Coordenação de Apefeiçoamento de Pessoal de Nível Superior

Investigating the genetic factors associated with sex determination system in Arapaima gigas (Pirarucu) using K-mer-based genome analysis.

17 Apr
1:00pm
P222

Renata Lilian Dantas Cavalcante[1]UFRN, J. Miguel Ortega[2] and Tetsu Sakamoto[1]
[1]UFRN; [2]UFMG

Arapaima gigas, popularly known as Pirarucu, is one of the largest freshwater bony fish in the world. It belongs to the family Arapaimidae, order Osteoglossiformes and has the Amazon Basin as its natural habitat. Adult specimens may weigh around 200 kilograms and measure about 3 meters. Due to its large size and its flesh containing low fat and low bone, along with its physiological characteristic of emerging to the surface at intervals of 15 minutes to assimilate oxygen, breeding of Pirarucu has been increasingly motivated, either with research purpose to better understand the specificities of the species, or to explore its economic potential. A problem related to its fishery exploitation is that the genetic mechanisms that control the sexual differentiation in A. gigas are not known. Sexual maturation of A. gigas occurs around the third to fifth year of life, and sexual dimorphism is not a strong characteristic of the species. For a more sustainable management, it is of paramount importance to seek an effective and non-invasive method to sexually differentiate juvenile individuals. For this, the establishment of a molecular genetic marker related to sexual differentiation would be an advantageous tool. Previous analyses comparing the A. gigas genomes of both sexes could not find genes or genomic regions that are associated with the sex-determination system of these individuals. In this study, we proposed to count the k-mers of the reads resulting from the sequencing of this species to identify regions in excess or missing in one of the sexes. For this, we used genomic data from four adult representatives of Arapaima gigas, two males (M1 and M2) and two females (F1 and F2), which in turn were submitted to Alignment and Assembly Free (AAF, v. 20171001), whose premise is to assemble a phylogenetic distribution based on the k-mers count that individuals share. In this analysis, the genome of Scleropages formosus, which is a species also belonging to the order Osteoglossiformes, was added as an outgroup. The k-mers counting of all samples generated by the AAF were then processed by in-house scripts to filter low-count k-mers, normalize (quantile normalization), transform (log2 and z-score transformation) and compare the data. The comparison between k-mers counts obtained among pairs of individuals of opposite or same sex was made by calculating the difference and the ratio between their counts. K-mer comparisons which exhibited extreme values were pooled, clustered using CD-HIT (v. 4.7), and analyzed individually. The CD-HIT analysis resulted in the formation of 1129 clusters of k-mers, of which 103 presented exclusively comparisons between opposite sexes and different individuals as well. In these initial exploratory studies, we have noticed the existence of k-mers over or underrepresented in one of the sexes, indicating differences in the genetic composition

between males and females of this species. These k-mers could be promissor candidates of molecular marker for sex differentiation in Pirarucu, although analyses with more samples are needed. Furthermore, k-mer-based methods for comparative genomic analysis demonstrated to be an interesting strategy to assist us on the unraveling of the sex determination system in Pirarucu.

In silico identification of candidate vaccine epitopes on Mycobacterium avium subspecies hominissuis strains: an investigation of the surfaceome

17 Apr
1:00pm
P223

Tayná da Silva Fiúza, Gustavo Antônio de Souza and João Paulo Matos Santos Lima
UFRN

Nontuberculous mycobacteria (NTM) are defined as environmental mycobacteria other than Mycobacterium leprae and Mycobacterium tuberculosis. In recent years the number of NTM infections has increased and Mycobacterium avium complex (MAC) is responsible for most of these infections. This complex is commonly described as an association of Mycobacterium intracellulare and Mycobacterium avium, where the latter species comprises four named subspecies with a range of infection capabilities and host species, causing opportunistic infections or aggressive pathological processes. Mycobacterium avium subsp. hominissuis infects humans and pigs, while some other subspecies infect cattle. To this day, MAC infection is controlled with multidrug regimens including macrolides, which doesn't seem to work as effectively and durably as it does for tuberculosis. The identification of specific molecular targets for controlling and removing NTM infections is an important and challenging task to which surface membrane proteins may be the answer due to the fact they are more easily accessible by the host immune system and as targets for drug therapy. Biochemical characterization of membrane proteins is often difficult due to the proteins hydrophobic nature, and even though specific protocols had been developed there is yet much to be perfected in those. We propose to use in silico surface membrane protein prediction followed by identification of core proteins, that is, proteins common to all strains, and then MHC binding affinity prediction as to to identify surface proteins that can trigger immune responses and are common to all strains of interest. To achieve that, we used amino acid sequences obtained from the NCBI database for all M. avium hominissuis strains with complete genomes sequenced. There were 201 genome correspondences for Mycobacterium avium on the Gene Assembly Summary of NCBI as of December, 2018. From these, only 18 had a Complete Genome status and only 7 of them were subspecies hominissuis strains. Therefore we worked with 7 strains and their proteomes in this project. Those sequences were submitted to alpha-helices transmembrane domains prediction by TMHMM, a machine-learning based tool. Sequences containing one or more TM alpha-helices were selected, as long as the helices comprised 18 or more amino acids and at least one helix occurred after the 60th amino acid. The TMHMM membrane alpha-helix prediction reduced the dataset size in from 32993 to 3429 proteins and comparison of all proteins revealed the majority are present in all seven strains, as it is expected for such close organisms. To define core proteins, the platform for Comparative Microbial Genomics (CMG) Biotools was elected for its efficiency and ease of use. There are 540 groups of homologous proteins (clusters) amongst the strains,

but only 393 of those represent core proteins. Next, immunological analysis was carried out using recommended methods available at the Immune Epitope Database and Analysis Resource (IEDB). For each protein, the immunogenetic prediction profile lists all different epitope-MHC allele combinations with their respective affinity scores. We defined a protein's Immunogenetic Score as the mean of the top 5% scoring epitopes in its sequence and classified a protein as Highly Immunogenic when its overall immunogenetic score figured amongst the top 5% in a given group. We found eight candidate proteins when filtering only core surfaceome clusters whose affinity for both MHC Class I and MHC Class II were amongst the top 5% highest scores. After identifying those clusters, we investigated whether the epitopes responsible for such high scores were reproducibly found in each of its protein members. We observed that cluster 315 is an epitope dense molecule in all strains, which makes it a good candidate for vaccine development. Other clusters show poorer results according to this metric, such as cluster 151 with few core epitopes, i.e., they were previously defined as highly immunogenic based on the score of independent peptides from each cluster member. In conclusion, we developed a protocol able to define core proteins in a proteomic dataset with the best immunogenic potential to be further tested/validated as a vaccine candidate. The protocol followed in this work will also be mostly automated as to allow other similar studies in pathogenic bacteria.

17 Apr
1:00pm
P224

Deep learning–based histological image recognition in colorectal cancer

Wanderson Gonçalves, Marcelo dos Santos, Fábio Lobato,Ândrea Ribeiro-dos-Santos and Gilderlanio S. Araújo

Universidade Federal do Pará

Histology has provided important predictive phenotypic information on the process and progression in cancer by identifying patterns of histological characteristics, like cell density, mitotic activity, nuclear atypia, and tissue architecture, and is therefore an essential tool for diagnosis and prognosis of the disease. Deep convolutional neural networks (CNNs) have been presented as a powerful tool in the analysis of images, with promising results in several study applications. The predictive ability in an image database makes it a potential instrument in the analysis of medical images. The present study aimed to evaluate the ability of deep learning on classification of histological images from human colorectal adenocarcinomas. For this purpose, a public database from the Kaggle platform was used, composing the database that fed the classifier training and testing process. Histological images samples are fully anonymized and it were extracted from formalin-fixed paraffin-embedded human colorectal adenocarcinomas (primary tumors) from pathology archive (Institute of Pathology, University Medical Center Mannheim, Heidelberg University, Mannheim, Germany). The database consisted of 5000 histological images of 150 x 150 px each (74 x 74 μm) for analysis, divided into eight categories of 625 images each (tumor, stroma, complex, lymphoid, debris, mucosa, adipose and empty). All images are RGB, 0.495 μm per pixel, scanned with an Aperio ScanScope (Aperio / Leica biosystems), magnification 20x. From the total images, 20% (1000 images), 125 of each category, were chosen randomly for the validation of the network. The deep convolutional neural networks were compiled with the Keras neural networks and the TensorFlow libraries through the python 3.6 programming language. The performance of CNN model was considered satisfactory, regarding the values of validation accuracy approximately equal to 80%. The results indicated that deep learning can be used to classify medical images of tissues with marked precision evidencing their potential and application in the analysis of histological images of colorectal cancer.

Evolution of the human neurotransmission network

17 Apr
1:00pm
P225

Danilo Oliveira Imparato[1],Lucas Henriques Viscardi[2],Maria Cátira Bortolini[2] and Rodrigo Juliani Siqueira Dalmolin[1]

[1]Bioinformatics Multidisciplinary Environment (BioME), UFRN, Natal, RN;[2]Departamento de Genética, UFRGS, Porto Alegre, RS

The evolutionary origins of behavior are a main theme in biology and its mechanisms are largely underlied by synaptic neurotransmission. It is understood that the last common ancestor of all synapses, the so-called ursynapse, emerged just before cnidarians and that the development of nervous systems followed shortly after. Orthologs to human synaptic elements are present in multiple organisms destitute of nervous systems and play role in ancient cellular processes. Such circumstances contribute to raise the question of how the interactions among these elements came to be and to what extent does it relate to our understanding of the evolution and complexity of nervous systems. In this work, we analyzed synapse-related genes in the context of their evolution, protein network an exclusivity to neural roles. Using the KEGG pathway database, we selected human genes present in the five main synaptic pathways: GABAergic, glutamatergic, serotonergic, dopaminergic and cholinergic. STRING database was used to access the genes protein-protein interaction network and to obtain orthology annotation for the available 238 eukaryote species as well as a species tree further manually curated. R package geneplast was used to infer orthologous groups emergence in the species tree according to orthologs presence/absence in extant taxa. We evaluated genes exclusivity to neural roles by means of both their number of pathways and their expression in nervous system tissues (Expression Atlas). Our results outline the distribution of neurotransmitter systems and genes synaptic roles as taxa diverge. We propose that human synapses genetic archetype was already established since our divergence with fishes and that the emergence of ionotropic receptors in the cnidarian last common ancestor was the driving force for the development of nervous systems as we know today.

Author Index