Proceedings X-Meeting 2019

Editor: AB^3C

October 2019

Conference Program

1 — Organizing Committee

AB3C President : Ney Lemke (UNESP)

AB3C Vice President : Marcelo Brandão (Unicamp)

AB3C Secretaries :

- Fabrício Martins Lopes (UTFPR)
- Vasco Ariston de Carvalho Azevedo (UFMG)

AB3C Financial Department :

- Nicole Scherer (INCA)
- Arthur Gruber (USP)

Poster Session Organizers :

- Alexandre Paschoal (UTFPR)
- André Yoshiaki Kashiwabara (USP)

Highlight Track Organizer :

- Arthur Gruber (USP)

2 — Introduction

The Brazilian Association of Bioinformatics and Computational Biology (AB3C) is a scientific society funded in July 12th 2004. Since its creation, AB3C has been responsible for the annual conference entitled ”X-Meeting” which is the main Bioinformatics and Computation Biology event in Brazil. This year its 13th edition occurred in Campos do Jordão from October 30th to November 1st .

Bioinformatics is now a strategic area for Brazil and all Latin America and, therefore, it is also strategic to the development of Science, Technology and Economy. The X-Meeting is a Brazilian event with international reach which has an average of 200 participants. The Conference is an opportunity for students, researchers and companies to interact and difuse knowledge. The AB3C has been a pioneer society in the field of Bioinformatics in Brazil and we have a history of ten past very productive meetings.

3 — Database and Software Development

,

Machado: a genomic data integration framework for Chado developed with Django

Mauricio de Alvarenga Mudadu, Adhemar Zerlotini Neto

Embrapa Informática Agropecuária

Abstract

Technological advances in biological research has led to a data deluge which have great impact in agriculture, especially in plant and animal breeding. In this regard, genome projects and multiomics experiments generate huge volumes of data that must be stored, mined and transformed into useful knowledge. Furthermore, all this information is supposed to be accessible and, if possible, browseable afterwards. Computational biologists have been dealing with this scenario for over a decade and have been implementing software libraries, toolkits, platforms, and databases to succeed in this matter. Although public wide databases exist, research groups still struggle to store and analyze data with local resources and expertise. The GMOD, or Generic Model Organism Database project, is currently the initiative that made the most advance in producing a "collection of open source software tools for managing, visualizing, storing and disseminating genetic and genomic data". Its biological relational database schema known as Chado is widely adopted and many softwares are able to connect to it. Such softwares usually contain a set of scripts to preprocess the data files in order to provide visualization and search capabilities, but in fact they can only operate independently. The Embrapa's Bioinformatics Multi-user Laboratory have been developing an open source software known as machado, that has a Django model to connect to Chado, thus avoiding extra efforts to make data compatible to the database schema. The machado software has several data loading tools for genomic and transcriptomic data and also for annotation results for tools such as BLAST, InterproScan, OrthoMCL and lsTrap. Machado has also an API to connect to Jbrowse which can be easily setup. A web browsing visualisation tool is implemented using the Apache2 server alongside with Django WSGI library. The haystack software integrated with the elasticsearch engine was used to create an index for querying the data using keywords, identification and annotation entries. Caching is also enabled for fast data retrieving. This project aims to contribute to the research community by producing a modern object-relational framework to store, integrate, query, and visualize all the major genomics data types. Such endeavour takes advantage of the latest Python modules to produce an effective open source resource and facilitate the identification of specific biological components related to economic traits in agriculture. Machado is available at GitHub: https://github.com/lmb-embrapa/machado .

Funding: Embrapa

,,,

MONET – SMARTPHONE APPLICATION FOR VIEWING MOLECULAR INTERACTIONS IN VIRTUAL REALITY ENVIRONMENT

Jorge Henrique Faine Monteiro, José Rafael Pilan, Agnes Alessandra Sekijima Takeda, José Luiz Rybarczyk Filho

Instituto de Biociências de Botucatu - UNESP

Abstract

Metabolic pathways correspond to processes that determine physiological and biochemical properties of a cell. These can be represented as interaction networks between proteins, metabolites and other molecules. Most current network visualization applications do not allow a three-dimensional view, displaying a planar network representation (two dimensions). In the present work, we developed a smartphone application MoNet (Molecular Network) that allows visualization of networks in three dimensions, enabling to the user an immersion in Virtual Reality (VR) environment, and providing a better experience of network analysis. The application was developed using Game Engine Unity3D, which was chosen because the ease to create three dimensional environments, the possibility to export to various smartphone platforms and effortless integration with Virtual Reality technologies. The C# was also used to query the information from the protein-protein interaction STRING database. The mobile user interface allows the user to select the organism, the proteins to be prospected, the interaction sources, the confidence score and the number of interaction limit per protein. The STRING returns a datafile containing the network (flat format) and protein descriptions. These network information are converted to a 3D representation by an algorithm adapted from the Force-directed graph drawing methodology that distributes proteins in the virtual space, allowing better visualization of the network. The user can place the smartphone in the cardboard holder to view the network in VR and even focus on a specific protein to visualize its description as well as the neighbors. MoNet also works without the VR environment, allowing zoom and rotation of the network. This first version of the application can be used to teach system biology to undergraduate and graduate students. For the next versions, we will include more databases like STITCH, that provides small molecule interaction and proteins, and STRING viruses, that provide protein-protein interactions between viruses and host.

Funding: CNPq

,

INN - Involuntary Learning Neural Network

Aline Rodrigheri Ioste, Alan Durham

Universidade de São Paulo

Abstract

For many scientific problems, the sharing of information between different groups in collaborative networks is essential for the combined analysis of the data. However, in many cases there are many ethical and legal problems in sharing the original data, in particular for medical research. This problem is even more restrictive across international borders as privacy and security legislations vary strongly from country to country. This article presents a new methodology to transfer and share information on sensitive data used to train neural networks. Using only the parameters of separately trained Neural Networks of the same architecture, this approach produces a "combined" network that has a performance similar to that of a Neural Network trained with the combined training sets of the original Networks. By transferring only the parameters of the trained networks, the approach maintains the secrecy of the original data. We compared the results of the approach using 5-fold cross validation obtaining, in the majority of cases, a correlation index higher than 0.9 between the Network obtained by our approach and that trained with the combined training sets. These initial experiments show the potential of our new approach, which can help the development of future systems for collaborative research even in fields with complex rules governing secrecy and privacy issues. In this poster we will present the basic algorithm being proposed, the current results applied to a simpler problems and discuss future developments.

Funding:

,,,

HT Atlas v1.0 database: redefining human and mouse housekeeping genes by mining massive RNA-seq datasets

Bidossessi Wilfried Hounkpe, Francine Chenou, Franciele de Lima, Erich Vinicius de Paula

Faculty of Medical Sciences, Unicamp

Abstract

Housekeeping (HK) genes are constitutively expressed genes that are required for the maintenance of basic cellular functions. Despite their importance in the calibration of gene expression, as well as the understanding of many genomic and evolutionary features, important discrepancies have been observed in studies that previously identified these genes. Here, we present Housekeeping Transcript Atlas (HT Atlas v1.0, www.housekeeping.unicamp.br) a web-based database which addresses some of the previously observed limitations in the identification of these genes, and offers a more accurate database of human and mouse HK genes and transcripts. The database was generated by mining massive human and mouse RNA-seq data sets from GTEx portal and ARCHS4 database. In total, 12, 482 and 507 high-quality RNA-seq data sets from 82 human non-disease tissues/cells and 15 healthy tissues/cells of C57BL/6 wild type mouse, were respectively included in our workflow. 2, 158 human transcripts from 2, 176 genes fulfilled our criteria and were refered as HK transcripts and HK genes. In the mouse database, 3, 024 HK transcripts from 3, 277 HK genes were identified. From the web interface, user can visualize the expression of those transcripts across tissues and download full lists of HK genes and transcripts. HT Atlas v1.0 also offers the most stable and suitable tissue selective reference transcripts for normalization of qPCR experiments. Some reference transcript-specific primers and predicted modifiers of gene expression for some of these HK transcripts are also proposed. All of these resources can be accessed and downloaded from any computer or small device web browsers. The database is a dockerized ShinyApp that can also be pulled from docker hub (bidossessi/ht_atlas_v1.0) in order to be locally deployed by the user.

Funding: FAPESP grants # 2016/14172-6, 2015/24666-3; CNPq Brazil. grant # 309317/2016

”

UNRAVELING MIRTRONS KNOWLEDGE WITH DATA MINING AND BIOINFORMATICS METHODS

Bruno Henrique Ribeiro da Fonseca, Douglas Silva Domingues, Alexandre R Paschoal

UTFPR

Abstract

In the current literature, there was no repository centralizing and organizing the mirtrons data available to the public. To fll this gap, we developed mirtronDB, the frst knowledge database dedicated to mirtrons, and it is available at http://mirtrondb.cp.utfpr.edu.br/. MirtronDB currently contains a total of 1, 407 precursors and 2, 426 mature sequences in 18 species (chordates, invertebrates, and plants). Our bioinformatics and data mining analysis highlighted that most studies on mirtrons were focused on H. sapiens and M. musculus. Consequently, we identifed more similarity results among chordates than in the other groups. This study provide initial mirtron characterization and can be used as a guide about mirtrons research.

Funding:

,,,,,

Text Mining for Biological Data: An Update from the Last Decade

Camilla Reginatto De Pierri, Diogo de Jesus Soares Machado, Bruno Thiago de Lima Nichio, Antonio Camilo da Silva Filho, Fabio de Oliveira Pedrosa, Roberto Tadeu Raittz

Federal University of Paraná

Abstract

Scientific literature is the basis of research in any field of study. Analysis of information from scientific texts is an important strategy to define the starting point and evaluate the state of the art in a given field of research, as well as assisting in the construction of hypotheses and interpretation of results. In the biological area, with the growing number of scientific texts deposited in public databases, the task of identifying relevant studies becomes complex and time consuming. To deal with this large amount of information, Text Mining (TM) approaches efficiently handle knowledge seeking. TM is a process that refers to the extraction of information found in texts. The advantage of TM techniques is the ability to improve bibliographic search, facilitate analysis and data storage, making the search process refined and accurate. Currently, there are series of TM tools that contemplate different methodologies with potential for study in the most varied biological scenarios. However, we noticed that most studies using TM as a research tool do not relate findings to a set of strategies, which in our view, may limit the discovery of knowledge. To assist researchers in choosing the best TM strategy, we conducted a literature review on the topic TM and the main tools available. The criteria for study selection were: 1) TM tools developed and / or implemented from 2009 to 2019; 2) Only tools available through scientific publication; 3) Only tools that the research was published in PubMed database. We have identified 41 TM tools with the most varied applications. These include MedlineRanker, DataShield, FACTA +, SAPIENTA, BioC and DISEASES with the highest number of citations, according to Google Scholar, Scopus and Web of Science. The methodology of the tools selected in this research involves processes of information retrieval, machine learning, natural language processing and computational and statistical language, focusing mainly on the study and identification of events related to genes and proteins. We found that text mining is not a simple keyword search in databases. Several automated processes and methods are required for the extraction of knowledge from texts. The use of one or more TM approaches is valuable for identifying relevant concepts and uncovering hidden knowledge in light of unexplored subjects.

Funding: CAPES, Fundação Araucária

”

Collscience: A Text Mining Web Service Tool for Extracting Knowledge from Scientific Texts

Diogo de Jesus Soares Machado, Camilla Reginatto De Pierri, Roberto Tadeu Raittz

Federal University of Paraná

Abstract

Manual curation of scientific texts requires effort and, depending on the research, involves a large number of researchers, which may not be sufficient to deal with the number of articles published in a given area of knowledge. In addition, manual curation tends to be affected by disagreement over interpretation, meaning different readers may interpret a specific text in different ways. Because of this, applications like Text Mining have become common in scientific research. With the emergence of the Knowledge Discovery from Text (KDT) definition, the mining process has gained notoriety due to the possibility of discovering important concepts in a variety of subjects. There are many text mining tools that are effective at the moment, but with the use of large amounts of text processing becomes very slow. Therefore, to address the difficulty of high computational effort in word processing, we propose Collscience (Collective Conscience), a text mining web service tool that integrates the best of information retrieval concepts with agility. The algorithm consists of a text mining pipeline: 1) compiling the texts to a format based on the representation of proteins in the FASTA format; 2) process the compiled texts in FASTA format using the SWeeP method, a methodology for vectoring and projecting data in FASTA format; 3) cluster using an improved version of k-means, predicting cluster numbers using a strategy that combines hierarchical clustering and Fuzzy logic; 4) define TF-IDF term scores, considering each text cluster as a document unit; 5) Put the terms in order of importance, using TF-IDF scores as a basis; 6) select TF-IDF scores for the most important words, with a user-defined total; 7) Perform hierarchical grouping, with the TF-IDF scores of the selected terms as a parameter; 8) Generate a dendrogram with the result of clustering. We also developed a graphical interface for Collscience, in web application format, using the languages HTML, CSS, JavaScript and PHP. Collscience is an algorithm that has a text set as input and performs mining, returning a dendrogram that shows the word correlation. The method is currently in the implementation phase. As future prospects, we plan to add more output options to the tool as well as improve user interaction with the machine learning process.

Funding: CAPES, Fundação Araucária

”

PapC: A web tool for paper clustering associated with PDB files

Daniel Viana, Wesley Paulino Fernandes Maciel, Raquel Melo Minardi

UFMG

Abstract

The literature review is the basis and first step required for any scientific study. Usually, the researcher searches in physical and digital scientific article repositories with the purpose of identifying content related to the researched subject. A search involves from the terms representing entities of interest as well as qualified relationships among them. Manly by the large volume of information available on the internet, this search returns vast data sets which makes the analysis process hard and toilsome. An example of this type of search and what motivated this study is the search for information related to a specific PDB file in a clear and precise way. When a new protein structure is published, it is deposited in the Protein Data Bank, and its structure is made available through the PDB files. With these files, it is possible to identify data related to some a publication. With this information, we can relate, through an API provided by NCBI to work with PubMed, all the papers relevant to the study of the respective structure. Hence, it can recover the related papers and all the articles that cited it. Therefore, this tool contains a database of relevant papers to each structure available at the PDB. The tool was created using Python, PHP, HTML, javascript, and CSS. To create the data visualization, we used the d3.js library. Preliminary results show that through a data visualization in graph form, it is possible to identify clusters of paper that can be related and relevant for a detailed literature review on the subject. Finally, it is also possible to identify collections of papers associated with several PDB ids that may indicate their relevance to the study of structures.

Funding: CAPES

”

Machine learning for predicting chemical-protein relations using graph embeddings

Daniel Viana, Raquel Melo Minardi, Adriano Alonso Veloso

UFMG

Abstract

The volume of biological data available in many repositories is vast and increases almost exponentially. Among this universe of data, we can highlight the relation between proteins and ligands. These relations may be the key to understanding biological processes, drug metabolism, drug design and repositioning, and industrial protein optimization. However, the large amount of data makes the process of analyzing and obtaining them manually a hard and toilsome process, hence making the use of in silico methods indispensable. One of the computational approaches that can be used in this case is graph modeling, where it is possible to represent proteins and ligands as nodes and the relationship between them as edges. So, this work aims to propose a new database of chemical compounds and genomic products graph-based from an unstructured corpus manually curated and to suggest a method capable of predicting a relationship between them through machine learning techniques. For this, we use the corpus chemical-protein interactions available on BioCreative VI Challenge. In order to predict relationships, we first use the Neighborhood Based Node Embeddings (NBNE) algorithm, an unsupervised method capable of generating node embeddings for graphs. Thus we produce a dataset containing the embeddings that represent nodes of the graph that contain a real relation, class one. For class zero, we generate false relationships randomly. To create prediction models, we use the machine learning algorithms: Decision Tree, Random Forest, and SVM as classifiers. Among the experiments performed, the method obtained the best result was SVM. Given the above, the base created is a way to provide and organize unstructured data of relationships between genomic products and chemical compounds and can be used as a query for possible relationships between them. Finally, the proposed method demonstrated to be efficient to suggest new relationships amid graph components through machine learning.

Funding: CAPES

,,,,,,,

AmpFLow: a containerized pipeline to assist in Reproducible and Replicable Microbiome research

David Aciole Barbosa, Fabiano Menegidio, Yara Natercia Lima Faustino de Maria, Rafael dos Santos Gonçalves, Marcos Vinicius Yano, Regina Costa de Oliveira, Daniela L. Jabes, Luiz R. Nunes

UMC

Abstract

The increasing number of studies aimed at evaluating microorganism populations is providing vast and detailed information about a large variety of environments. Such microbiome data represent one of the most promising approaches currently available in biological sciences, with possible applications in industry, agriculture and health. However, technical obstacles still must be overcome, before results from microbiome analyses can be fully incorporated into innovative technologies. In fact, scientists around the world claim that many fields of research are currently affected by a reproducibility challenge, due to difficulties in obtaining all details and resources necessary to reproduce the same experiment/analysis in different laboratories. Microbiome analyses seem to be particularly affected by such reproducibility crisis and the lack of proper standardization for the complex bioinformatics procedures, inherent to such analyses, seems to be one of the main reasons for such problems. A new trend to solve this issue is Docker, a system-independent technology which made possible the packaging of complete software environments by creating additional layers of operating system level virtualization abstraction bundled in the so called containers. In this sense, we present AmpFlow, a containerized pipeline designed to promote reproducible and replicable microbiome analyses. AmpFlow was built in Docker, using a set of simple, yet effective scripts, to deploy tools available from Qiime, which perform: (i) quality checking; (ii) pre-processing and (iii) processing of raw bacterial and fungal sequencing data, creating reproducible OTU tables that are ready to be used in different post-processing platforms, such as the online and R versions of Microbiome Analyst, as well as the Galaxy version of LEfSe.

Funding: FAPESP, CAPES, CNPQ

,,,

Goliath, a NGS web-based platform.

David Berl, Thiago Luiz Araujo Miller, Daniel T. Ohara, Pedro Alexandre Favoretto Galante

USP

Abstract

The development of Next-generation sequencing platforms (NGS) in the past decade created an accurate and cost-effective methodology with application in many areas ranging from basic research to individual patient care. Nowadays, it is trouble-free, fast and inexpensive to generate NGS data. However, this ongoing revolution is placing a significant demand for expertise in processing these large sequencing datasets produced by NGS platforms. It is not rare to see researchers with NGS data on hand and stuck in the step of processing these data. Several initiatives have been produced in order to simplify the NGS data processing, but most of them have pitfalls, such as incompleteness in terms of pipelines and/or difficulty of usage. We present Goliath, a web-accessible platform for NGS processing. At its core functionality, Goliath will provide support for both transcriptomics and genomics analysis. In its launch version, Goliath processes FASTQ format data from human RNA sequencing (RNA-seq) and produces both gene expression patterns and differential expression. With a receptive interface, multiple samples and their replicates can be uploaded and combined to create an assortment of comparisons. This is achieved using the latest reference transcriptome available (e.g., GENCODE, for humans) and some of the most commonly used algorithms to quantify gene expression (e.g., Kallisto, which uses an alignment-free approach), and DESeq2 for obtaining the set of differentially expressed genes. Goliath also uses the R environment to produce clever and customizable graphs. In essence, Goliath aims to aid researchers that can produce NGS data for elucidating relevant biological questions but have a tough time processing these NGS data by their own.

Funding: CNPq

,

Feature selection from data integration through analysis of Copy Number Variation (CNV) for genotype-phenotype association of complex diseases

Christian Reis Meneguin, David Correa Martins Jr

Universidade Federal do ABC - UFABC

Abstract

Complex diseases are usually consequences of intracellular and intercellular disorders in tissues and organs, being developed in a multifactorial way. Togetherwiththeproductionofafairlyhighvolumeofbiologicaldatagener performancesequencingtechniques, researchesinthisareanow involve integrative data analyses. In this context a computational framework was created to allow the integration of different types of biological data from the chromosomal location. This tool was used to study autism spectrum disorders (ASD), showing regions of Copy Number Variation - CNVs present only in samples of affected individuals, becoming a promising work to aid researches involving complex diseases.

Funding:

,,,

Accurate Identification of Hosts from Environmental Viruses Using Deep Learning Networks and High Level Features

Deyvid Amgarten, Bruno Iha, Aline Maria da Silva, João Carlos Setubal

Laboratório de Técnicas Especiais, Hospital Israelita Albert Einstein

Abstract

Microbial genomics has been experiencing expressive changes in the last decade, mostly due to improvements in environmental sampling and sequencing techniques generally known as metagenomics. Thousands of new complete virus genomes are made available every year, but experimental characterization has not kept pace. In particular, information about the host of viruses whose genomes have been sequenced is commonly lacking. This is one of the most essential information needed for cultivation, isolation and many other microbiological characterization techniques. Here we present a toolkit called vHULK (Viral Host Unveiling Kit), which predicts taxonomic and biological attributes of a virus' host. Our tool receives complete or high quality virus draft genomes as input, and provides as output: 1. Probability scores of predicted host genus and/or species; 2. Tables of cross-probabilities among possible hosts; and 3. information theory measurements and a rank of informative features about virus-host relationships. Our methodology is based on feature extraction of virus' genomes with known hosts leading to matrices containing thousands of features. After feature extraction and feature normalization, matrices were used to train a multilayer perceptron deep neural network classifier. Performance was measured by assessing validation and training curves, as well as by measurement in batch test sets. For a multiclass problem of 62 possible bacterial host genus, vHULK presented an average accuracy of 98% in the test set. When development is finished, vHULK will be freely available through user-friendly python scripts at a Github repository.

Funding: This work had the support from CAPES, CNPq and FAPESP research funding agencies

”

Bootstrap approach for multivariate survival analysis of cancer patients.

felipe rodolfo camargo dos santos, Gabriela Der Agopian Guardia, Pedro A F Galante

Instituto de Ensino e Pesquisa - Hospital Sírio Libanês

Abstract

Gene expression is an important factor correlated with survival of cancer patients. It has been shown that tumors harbour many synergistic and antagonistic interactions that impact disease progression. Given this complexity, the use of regression methods for survival analysis is a valuable resource for understanding the molecular profile of tumors and its impact on tumor progression, especially multivariate models, which take into account the important contribution of the context in which each variable is regarded. Despite their importance, the existing algorithms present some limitations, in particular for the analysis of high dimensional datasets, such as handling attribute/observation proportion issues (dimensionality problem) and control of attribute-attribute correlations. In order to achieve analysis convergence for high dimensional datasets we propose a bootstrap approach, making use of current regression methods that, in turn, use lasso and/or ridge parameters for the obtention of reliable regression coefficients in a bias-variance trade-off optimization. Our tool was applied in a real scenario for the investigation of survival differences between men and women patients with glioblastoma and other cancer types. These analyses revealed protein coding genes and long non-coding RNAs (lncRNAs) that may contribute to prognosis differences based on the gender of patients. Interestingly, these enriched lncRNAs are still poorly characterized and potentially subject to further investigations. Our results suggest that multivariate analysis combined with bootstrap algorithm improves prognostic prediction in comparison with commonly used methods that rely on univariate regression filters for lowering data dimensionality. In summary, we believe our method provides a more reliable and adjusted list of genes than current strategies because it takes into account gene-gene interactions.

Funding: Fapesp

,

Improving protein-based metagenomic reads classification

Giovanni Marques de Castro, Francisco Pereira Lobo

Universidade Federal de Minas Gerais

Abstract

The most used software for the classification of metagenomic reads have the option for the user to build a custom database, which is expected to be updated and encompassing as much taxa as possible. The proposal of KAIJU is that amino-acids are more conserved and using them increase the sensibility, especially from samples of extreme environments. However, the database proposed by them that includes the widest range of taxa is the nr_euk. For metagenomic studies that focuses on fungal organims, the results obtained can be much worse than using a nucleotide search, missing most of the reads and having an increased misclassification. This problem arises because fungal genomes are deposited in GenBank without their protein annotation, this is reflected in the source database used by KAIJU, the NR, not containing the peptides from those genomes. After downloading 2027 fungal genomes from GenBank that passed some filters, only 794 had their protein file available to download (*protein.faa.gz in the GenBank ftp of the genome). This means that 60% of the fungal genomes do not have their proteomes available. To generate a protein database that is broad and includes information from those fungal genomes, genomes from bacteria, viruses, archaea and from those downloaded fungal genomes were translated in their six frames. Due to RAM limitation, only peptide sequences with over 60 amino acids were kept and indexed using KAIJU. This is an strategy similar to tblastx, but much faster as it uses KAIJU. In a preliminar result with a metagenomic sample, CENTRIFUGE, a nucleotide based classifier, using a database with the same downloaded genomes, classified over 600.000 reads as Lasiodiplodia. On the other hand, KAIJU using the nr_euk classified only a bit over 5000 reads as Lasiodiplodia and over 150.000 reads as Diplodia, both Fungi in the Botryosphaeriaceae family. Nonetheless, when using the database of six frames translated genomes, KAIJU was able to classify around 600.000 reads as Lasiodiplodia, while almost no reads as Diplodia, close to the result returned from CENTRIFUGE. Lasiodiplodia is an example of a gender of Fungi without a single predicted proteome, moreover the single genome available have over 97% completeness when analyzing its quality with BUSCO, so it should be well assembled by this parameter. With this strategy to build a protein database, not only KAIJU, but other protein-based classifiers could use it, as the NR lacks over half of the potential proteins from Fungi in which the genomic information is already public available.

Funding:

”

sideRETRO: structural variations intercurences discovery environment for retrocopies

José Leonel Lemos Buzzo, Thiago Luiz Araujo Miller, Pedro Alexandre Favoretto Galante

USP

Abstract

The understanding of Transposable elements' mechanisms are gaining a rising focus on actual researches as key features of genomic structures and the impacts of their dynamics are directly associated with many pathological scenarios. It has been clearly shown the protagonic role exerted, for example, by somatic retrotransposon insertions in tumorigenenic cases, whether disrupting promoters of critical protein coding genes, or creating new splicing sites and isoforms, or even becoming expressed. Therefore, accurate detection methods ought to be settled down for them, methods by which whole genomic assessments of the copy number variations of these transposable elements would be feasible. However, some quantitative difficulties lies on this task concerning the ambiguity on mapping new inserts: actual aligners frequently report their reads as belonging to the parenthal gene in the reference genome. Corroborating this fact, literature shows only a few examples of bioinformatics tools available to this errand and, even these, cannot ensure their accuracy because of a lack in false positive statistical controls. So, to get around this problem, a candidate algorithm would need to (1) distinguish ambiguous read mappings from their parenthals and other fixed copies of it, based on an annotated gene list; and (2) robustly learn to discern the false positive cases using some simulation approach. Now focusing on retrocopies, which are new insertions of processed protein coding genes' mRNAs made by LINE retrotransposon machinery, we present sideRETRO, a computational bioinformatis tool for polymorphic and somatic retrocopies discovery, in a genomic landscape. Provided with an accuracy tuning simulation method suited for this task and a ambiguity solving engine based on discordant reads mappings, our tool was used to search for retrocopies on ten whole genomes sequenced from healthy individuals with more than eighty years old. It was chosen only healty individuals because of the need to compose a non tumorigenic retrocopies profile when comparing to future pathological samples. So, assessed by a previous simulation batch for false positive filtering, our algorithm reached a 0.87 accuracy rate on a 5-fold cross validation. And, when used on the ten whole genomes with high coverage (40x), it discovered a mean of seventeen retrocopies per genome, which were further identified when polymorphic or not and when cancer related or not.

Funding: CNPq

,,,,,

JMSA2: Java Mass Sectrometry Analyzer

Bruno Henrique Meyer, Malton William Machado Cunico, Dieval Guizelini, Emanuel Maltempi de Souza, Fabio de Oliveira Pedrosa, Leonardo Magalhães Cruz

Federal University of Parana

Abstract

The use of MALDI-TOF mass spectrometry allows microorganism identification by generating mass spectra representing a characteristic profile of signals from ionized whole cell peptides or cell extracts. The microorganism identification by means of mass spectrometry is a recent technique, applied to different types of samples (i.e., clinic or environmental), with many advantages compared to classical approaches (e.g., amplification and sequence of genetic markes, such as 16S rRNA gene). It is time and cost effective. Comparing of mass spectra obtained from unknown microorganisms with a database of mass spectra for known microorganisms allows their identification. However, the spectra generated are complex data, being its interpretation and analysis difficult. Further, among few alternatives of software, there are proprietary code ones with limited environmental representative databases. The Java Mass Spectrometry Analyzer (JMSA) has been developed in open source code by the Nucleus of Nitrogen Fixation at Federal University of Paraná (UFPR) that facilitate the visualization, manipulation, creation of databases, and comparison of mass spectra for the purpose of microorganism identification, as well as include descriptive sample data. Here, we present JMSA version 2, developed in Java (platform independent), with the following main characteristics: i) clustering algorithm; ii) export results in many different file formats; iii) build superspectra that enhance comparison and identification; iv) creation of spectra database; v) pairwise-spectra comparison; vi) spectra database search tool for microorganism identification.

Funding: Supported by INCT-FBN, CNPq, and CAPES

,

An integrated approach to building and applying profile HMMs for sequence detection in genomic and metagenomic data

Liliane Santana Oliveira Kashiwabara, Arthur Gruber

USP

Abstract

In this work, we report the development of an integrated approach for the construction of profile HMMs and their use in a series of applications using genomic and metagenomic data. To construct profile HMMs, we developed TABAJARA, a program for the rational design of profile HMMs using multiple sequence alignment (MSA) files. By optimizing a position-specific information score in a sliding window along the length of an MSA, TABAJARA automatically identifies the most informative sequence motifs that are either (1) conserved across all sequences or (2) discriminative for two or more specific groups of sequences, and then constructs profile HMMs from these alignment blocks. The generated models can be used to screen genomic or metagenomic sequencing data with HMM-Prospector program, a tool that performs similarity searches and quantifies the results according to score or e-value thresholds for each tested profile HMM. Models displaying the most relevant results can be used as seeds by GenSeed-HMM, another tool developed by our group to perform seed-driven progressive assembly using unassembled sequencing data. Finally, profile HMMs can also be used by the program e-Finder to identify and extract multigenic elements, starting from assembled genomes or metagenomes. Potential applications include the detection of proviruses, mobile genetic elements or any other set of specific genes present in a specific syntenic context, such as operons. We obtained a successful set of results using this toolbox for studying casposons, a family of self-synthesizing mobile elements that are found in archaeal and bacterial genomes and that gave rise to the currently known CRISPR elements. First, we designed casposon-specific profile HMMs constructed from endonuclease Cas1 and DNA polymerase B sequences. These models were used to screen several unassembled metagenomic datasets with HMM-Prospector. The positive sets were submitted to progressive assembly with GenSeed-HMM program, using the profile HMMs as seeds, resulting in the reconstruction of casposon-specific sequences. Also, the same models were used with e-Finder to detect and retrieve casposons sequences from assembled bacterial and archaeal genomes of the PATRIC database. A total of 138 elements derived from 105 distinct genomes were detected. This number of elements is approximately three times higher than the number of casposons reported in the literature. In both cases, phylogenetic analyses confirmed the correct taxonomic assignment of the positive sequences.

Funding: CAPES

”

Plant Co-expression Annotation Tool: a tool to identify targets for proof of concept in Genetically Modified crop breeding pipelines

Marcos José Andrade Viana, Adhemar Zerlotini Neto, Mauricio de Alvarenga Mudadu

UFMG/EMBRAPA

Abstract

The development of Genetically Modified crops (GM) includes the discovery of candidate genes through bioinformatics analysis using genome data, biological pathways, gene expression, and others. Proteins of unknown function (PUFs) are interesting targets for proof of concept in GM crops breeding pipelines. Many PUFs are species specific and may participate in important biological pathways for organism survival. One way of inferring the function of PUFs (e.g.: relating them to factors of interest, like abiotic stresses), is through orthology and gene expression correlations by using co-expression networks. The purpose of this work is to characterize PUFs to be used in GM crops pipelines using orthology, co-expression networks and other tools. To date, we have downloaded, analyzed, and processed genomic data of 53 organisms from Phytozome, as well as the genome of the resurgent Boea Hygrometrica plant from NCBI, totaling 1, 862, 010 genes, 2, 332, 974 mRNA and 2, 332, 974 proteins. Diamond blast, against the NCBI's nr database, and InterproScan were used to discover 72, 266 PUFs for all organisms. PUFs were found by selecting proteins that have no matches within both diamond and interproscan searches. To construct the co-expression networks, RNA-seq data related to abiotic stress, like heat, drought, dehydration and osmotic stress were downloaded from the GEO / NCBI database: 16 samples from Glycine max, 14 from Zea mays 33 from Arabidopsis thaliana and 28 from Oryza sativa. This data was used to construct co-expression networks and clusters of transcripts with correlated expression using the LSTrAP software. Orthology was used to annotate the PUFs. In this regard, we have constructed 164, 267 orthologous groups using OrthoMCL software with the proteins from the 53 genomes. All the data obtained were stored in a database provided by the machado software (https://github.com/lmb-embrapa/machado). A web interface named "Plant Co-expression Annotation Tool" is under development to provide queries to mine PUFs from all 53 plant species. The tool provides analysis such as comparative functional annotation searches, expression values, biological pathways and ontologies. A search example was performed using the Oropetium thomaeum genome, a plant that is resistant to ressecation, and we found 10 PUFs correlated with abiotic stresses through orthology and co-expression networks. In summary, we believe our tool can be valuable for finding interesting targets to be used as proof of concept in GM crop breeding pipelines.

Funding: EMBRAPA/UFMG

,

PIMBA: A PIpeline for MetaBarcoding Analysis that allows the use of a personalized reference database.

Renato Renison Moreira Oliveira, Guilherme Oliveira

Universidade Federal de Minas Gerais

Abstract

DNA metabarcoding is an emerging field in biodiversity analysis. The increase in data generation thanks to Next-Generation Sequencing caused the development of new DNA metabarcoding techniques. Mothur, Qiime, Obitools, and mBRAVE are currently the most used metabarcoding analysis tools. The only problem with those pipelines is that they do not allow the researcher to use a personalized database. For example, Mothur is only useful when analyzing 16S data. Qiime (and even its updated version, Qiime2) is optimized to analyze metabarcoding data from 16S, 18S and fungal ITS marker genes, using Greengenes, SILVA and UNITE databases, respectively. If the researcher is interested in using the Qiime pipeline to analyze data sequenced from different marker genes, such as COI or Plant ITS, Qiime does not give support to adapt its pipeline for such personalized use. Obitools is optimized to analyze data from 16S (SILVA and PR2). Obitools also allows the use of the NCBI database for taxonomic assignment. mBRAVE is optimized to use only the BOLD database as a reference, allowing the researcher to use a personalized database only after BOLD submission. Here we present PIMBA, a pipeline for metabarcoding analysis which adapts the Qiime pipeline in order to allow the researcher to additionally use a personalized or the NCBI databases. PIMBA also implements all the features provided by the other metabarcoding tools such as reference databases for 16S, 18S and Fungal ITS. PIMBA performs the NCBI taxonomic assignment by blasting the resulting OTUs against the NCBI nt database and using taxdump, it retrieves the full taxonomic information, creating all the files that Qiime needs to perform its final analysis. PIMBA also allows the researcher to analyze datasets from marker genes without a published reference database (COI, Plant ITS, rbcL, matK, etc.), where it is only needed a fasta file with the full taxonomy written in the header of the sequences. We used PIMBA to analyze two metabarcoding datasets (Plant ITS and Invertebrates COI). PIMBA was able to generate all the outputs needed to infer results and additional metabarcoding analysis, such as alfa and beta diversity analyses, PCoA and taxonomic bar plots. Next steps include benchmarking PIMBA against the most used metabarcode tools. We will also create a docker image, making PIMBA easily deployable by the user.

Funding: 372439/2019-5, CNPq

,,,

Knowledge Management in Genomics: The Role of Data Provenance

Vinicius Werneck Salazar, Kary Ann del Carmen Ocaña Gautherot, Fabiano Lopes Thompson, Marta Mattoso

LNCC, National Laboratory of Scientific Computing, Petrópolis, Brazil

Abstract

Genomics as a discipline has grown considerably in recent years and its methods have proven to be helpful for solving diverse problems across various domains of the life sciences. It has been constantly driven by data life cycles, with scientists relying on public, curated data repositories to perform their own experiments. The consequences of this are that the issue of managing data correctly and efficiently have been central to the field, and with the exponential growth in publicly available data, this has turned into a "big data issue" which concerns individual scientists, research groups, institutions, and international consortia. We discuss how knowledge management (KM) can address some of the big data issues in genomics. The KM approach has proven to be valuable in case studies, due to its systematic process of defining workflows that improves efficiency and reproducibility of experiments. However, combining a KM workflow with genomics has challenges related to the genomics data life cycle like provenance. Tracking the provenance of data in genomics is becoming part of its data life cycle, but provenance data is not part of the KM workflow. Effective data management, good practices in scientific computing, FAIR data sharing, and collaborative work between experts in different disciplines can all be facilitated by provenance tracking. Provenance tracking is a key aspect in establishing the data-information-knowledge cycle, which can be used as a framework for planning, executing scientific and monitoring experiments. Because provenance is information providing context to data, it plays a fundamental role in generating and sharing knowledge at the end of the cycle. Provenance tracking needs data capture and storage, which may affect the performance of the genomics workflow. Our proposal is focused on bioinformatics workflows and how they should be planned from a KM provenance-based perspective. We show how to adopt a KM provenance-based perspective of experiments in genomics with negligible computational costs by adopting a dataflow analysis system, in this case, the DfAnalyzer tool, which adopts the W3C PROV standard. By coupling DfAnalyzer with a genomic bioinformatics pipeline, it is possible to capture provenance data which when queried generates information about trade-offs, for example, of performance and sensivity, providing the research with an enriched knowledge of his analysis. The different DfAnalyzer components align with those of knowledge management: the Provenance Data Extractor, Raw Data Extractor and Raw Data Index act on data capture and storage, the Query Interface and Dataflow Viewer on information summarizing and analysing, allowing a synthesis of knowledge around the experiment which can support decision making. This is particularly useful when scaling experiments with a large number of samples or parameter sweeping. For demonstration, we executed DfAnalyzer with a bacterial genome annotation and phylogenetic analysis workflow, using DfA, the DfAnalyzer Python library. We considered the steps of data and metadata collection, sequence statistics, gene call predictions, annotation with the NCBI COG database, and pairwise Average Nucleotide Identity between genomes, integrating each step to the Dfa library calls with a custom Python package. In our workflow case study we show the benefits of monitoring, querying and reuse of methods

,

Stochastic Models alongside Deep Learning Methods for Gene Prediction

Waldir Edison Farfan Caro, Alan Durham

Universidade de São Paulo

Abstract

Sequence labeling is the task of, given an observed sequence, determine the best label for each element according to a set of predefined categories. It has applications in many research areas where the detailed understanding of the sequence is mandatory, such as bioinformatics, natural language processing and computer vision. To address this problem, stochastic methods were developed such as Hidden Markov Models (HHMs) and Conditional Random Fields (CRFs) due to their ability to model the relationship of the members of the sequences. Lastly, Deep Neural Networks (DNNs) have been used satisfactorily for sequence labeling problem since they are suitable to automatically learn complex feature representation from data, which is an advantage to models designed by hand. Since CRFs have good achievements in modeling and DNNs have a high capacity in learning the representation. In this work we propose the use of mixed methods of CRFs and DNNs capable of improve the task of sequence labeling, in special the problem of gene prediction, by giving suitable models for the sequences with a rich representation of their features. For this porpouse, we use the ToPS framework, as a source of efficient implementation of Markov Models and the recently integration of CRFs; and the PyTorch framework which offers optimized tensor implementation for deep learning methods. Both frameworks follow an object oriented approach giving the chance to a better blending of the techniques. Thus, for gene prediction we gain better feature representation of genome sequences beside rich modeling of their structure, even for complex organisms.

Funding:

4 — Genes and Genomics

”

Regulatory elements of carbon metabolism in sugarcane

Alícia L.de Melo, Alan Durham, Glaucia Souza Mendes

Abstract

The increase in the access to renewable energy sources contributes to the global efforts to reduce greenhouse gas emissions. Among these sources, we highlight the cellulosic bioethanol technology, which is generated from plant parts, mainly sugarcane. Knowledge of the components of secondary cell wall metabolism regulation networks will allow the construction of biotechnological tools for the development of energy cane, which has more fiber and biomass, by changing the carbon partition of sugar to biomass. Synthetic promoters are a potential tool for the creation of technology that can be applied in the production of transgenic plants with the desired biomass characteristics or by the use of gene editing methodologies such as CRISPR-CAS. Thus, the characterization of the architecture of the promoters of the target genes involved in the carbon metabolism regulatory networks is essential to provide tools for transgene technology and gene editing. In the present work, a gene from the sugarcane SP80-3280 genome involved in the starch and sucrose metabolic pathway was selected for the initial testing of the analysis methodologies. An important aspect of motif discovery is the delimitation of the sequences that will be used to search for motifs. This delimitation was based on the location of the transcription start site, which was determined by three different approaches. Only one of these approaches yielded positive results for the detection of representative motifs. Based on that, it was possible to identify six potential transcription factor binding sites for this gene. The next steps involve the comparison of these detected motifs to the ones already described in the literature, as well as applying the methodology to a great number of genes at the same time.

Funding: Fundação de Amparo à Pesquisa do Estado de São Paulo

”

A bioinformatics approach to cancer vaccines prioritization based cancer-testis antigens in melanoma

Andre Fonseca, Ana Carolina Miranda Fernandes Coêlho, Sandro Jose de Souza

USP

Abstract

Peptide-based vaccines are a promising approach to cancer immunotherapy. Nowadays, several clinical trials using peptide-based vaccines have been carried in distinct tumor types, such as colorectal, lung cancer, and melanoma. Traditionally, these vaccines are synthesized based on tumor-associated antigens, such as cancer-testis antigens (CTAs). The CTAs are a large family of tumor-associated antigens with expression restricted in normal tissues (testis) and expressed in a broad number of tumor types. In addition, CTAs have been pointing out as important candidates due to low toxicity and high specificity. However, several attempts to used CTAs as vaccines were considered a failure, most likely due to a wrong candidate choice and/or absence of techniques to analyze the clinical risks. Consequently, it remains as an opened challenge, with two main bottlenecks: i) properly prioritization of novel candidates, and then ii) how to combine these candidates effectively. Here, we present a multiple CTA vaccine to approach melanoma patients. First, we analyzed a large collection of CTAs in the TCGA cohort, in which was filtered candidates based on individual contribution to good prognosis and high T cell CD8+ correlation. As a result, were selected seven candidates, including INSL3, HSF5, GTSF1L, FAM209A, GPR31, SIRPD, and HEATR9. Next, we clustered patients into distinct groups based on the number of co-expressed CTAs. In a total, four groups were obtained, named as Red (>5 co-expressed CTAs), Yellow (>2 to 4), Blue (>1 to 2), Grey (without expressed CTAs). Interestingly, the survival ratio has a significant improvement based on the number of co-expressed CTAs. Furthermore, it seems to be independent of other immune response drivers, such as mutation load. Finally, we evaluate the antigenic regions (epitopes) for each CTA protein. In short, the epitopes predictions were carried out using netMHC algorithms, considering a subset of high frequent HLA alleles based on population data. As a result, 40 and 27 high antigenicity regions were found to MHC I and II, respectively. Interestingly, a few peptides are associated with both CD4 and CD8 responses. In conclusion, these findings can be an interesting resource for a cancer vaccine protocol for melanoma.

Funding:

,,,,,,

Prediction, identification and characterization of genomic islands in Aeromonas spp.

Antonio Camilo da Silva Filho, Camilla Reginatto De Pierri, Diogo de Jesus Soares Machado, Roberto Tadeu Raittz, Jeroniza Nunes Marchaukoski, Cynthia Maria Telles Fadel-Picheth, Geraldo Picheth

Federal University of Paraná

Abstract

Aeromonas are pathogenic bacteria, mostly aquatic and with wide environmental distribution. These bacteria can cause infections in humans, varying in severity, for example in cases of uncomplicated acute gastroenteritis, where the disease does not pose serious health risks, and septicemia, which can be fatal. Virulence is the ability of a microorganism to cause disease in a host, and several features of the Aeromonas genome have been proven to be associated with virulence. Even though virulence characteristics are similar between species, there are not many studies that discuss the behavior of these genes in the Aeromonas genome in general. However, it is already known that one of the main mechanisms that bacteria use to share genetic material is the genomic islands (GIs). In this perspective, to understand and identify the main genes related to pathogenic potential and virulence mechanisms in Aeromonas strains, this study aims to characterize and compare the gene content and respective function of Gls in Aeromonas, as well as to analyze their GIs distribution between different species of this bacteria. For this, complete genomes of 58 Aeromonas were obtained from the NCBI Database. GI prediction was performed using IslandViewer4 software, which combines analyzes by comparative genomics and sequence composition; the two most used methodologies to identify these regions. To determine GI functions, we used the ARDB, CARD, NDARO (antibiotic resistance), Patric, VFDB, VICTORS (Virulence), DrugBank, TTD (drug targets) databases. Analysis of GI distribution among species was performed using the CD-HIT-2D cluster with a 70% self-score similarity. To evaluate the content of complete genomes in all Aeromonas studied, a phylogenetic tree was constructed using SWeeP method. The results showed that it is possible to determine the genetic diversity of these organisms and to characterize GIs according to their related functions and products using clustering and phylogeny. It was possible to identify the relationships between the origin of GIs (clinical / environmental / animal) in the total set of Aeromonas. Phylogenetic analysis complemented cluster analysis by showing that bacteria are misclassified, such as Aeromonas hydrophila YL17 and Aeromonas hydrophila 4AK4 strains that are not hydrophila species.

Funding: CAPES

,,,

Microbiomes of Velloziaceae from phosphorus-impoverished soils of the campos rupestres, a biodiversity hotspot

Antônio Pedro de Castello Branco da Rocha Camargo, Rafael Soares Correa de Souza, Paulo Arruda, Marcelo Falsarella Carazzolle

Universidade Estadual de Campinas

Abstract

The rocky, seasonally-dry and nutrient-impoverished soils of the Brazilian campos rupestres impose severe growth-limiting conditions on plants. Species of a dominant plant family, Velloziaceae, are highly specialized to low-nutrient conditions and seasonal water availability of this environment, where phosphorus (P) is the key limiting nutrient. Despite plant-microbe associations playing critical roles in stressful ecosystems, the contribution of these interactions in the campos rupestres remains poorly studied. We generated and investigated the first microbiome sequencing data of Velloziaceae spp. thriving in contrasting substrates of campos rupestres. We assessed the microbiomes of Vellozia epidendroides, which occupies shallow patches of soil, and Barbacenia macrantha, growing on exposed rocks. The prokaryotic and fungal profiles were assessed by rRNA barcode sequencing (16S V4 and ITS2) of epiphytic and endophytic compartments of roots, stems, leaves and surrounding soil/rocks. Through this data, we found that there is a large quantity of as-yet-unknown microorganisms thriving in the campos rupestres environment. When contrasting the microbiomes of the two plants, we observed major differences regarding the community composition, diversity and colonization profiles. Interestingly, we also noticed that there are several highly abundant microorganisms that associate with both V. epidendroides and B. Macrantha, suggesting a shared core microbiome in this environment. Shotgun sequencing of total DNA extracted from microbial samples of rhizosphere and substrate was performed to investigate the functional landscape of the campos rupestres microbiomes. The samples were individually assembled and annotated, generating a median number of 9, 907 noncoding genes and 2, 544, 611 protein-coding genes. The comparison between metabolic profiles of communities associated with the substrates and the rhizospheres of the two plants revealed major functional differences between the two microbiomes. We foresee that these data will contribute to decipher how the microbiome contributes to plant functioning in the campos rupestres, and to unravel new strategies for improved crop productivity in stressful environments.

Funding: FAPESP

””

Variants encompassing the Agouti signaling protein gene are associated with dilution of grey shades in Nellore (Bos indicus) cattle

Beatriz Batista Trigo, Marco Milanesi, Adam Taiti Harth Utsunomiya, José Fernando Garcia, Yuri T. Utsunomiya

1 Department of Support, Production and Animal Health, School of Veterinary Medicine, São Paulo State University (Unesp), Araçatuba/SP, Brazil; 2 International Atomic Energy Agency (IAEA) Collaborating Centre on Animal Genomics and Bioinformatics, Araçatuba/SP, Brazil; 3 Department of Preventive Veterinary Medicine and Animal Reproduction, School of Agricultural and Veterinarian Sciences, São Paulo State University (Unesp), Jaboticabal/SP, Brazil

Abstract

Nellore cattle (Bos indicus) are well known for their resilience to infectious diseases, survival in low input systems and tolerance to heat. Traits that are often speculated to contribute to heat tolerance are the skin pigmentation and the coat color. Nellore cattle posses dark-black skin, and although their coat color pattern can range widely from solid red to black-spotted, the Brazilian herds were massively selected for white coat. For this type of coat, whereas females are completely white, males can exhibit shades of dark gray in the head, neck, hump and knees. Curiously, some males are also completely white, and they can transmit this phenotype to their progeny in a dominant mode of inheritance. The present study focused on mapping the functional candidate gene responsible for the dominant white phenotype in Nellore cattle. Bulls were evaluated based on visual scores for gray shading in a subjective scale from 1 to 5 by six evaluators, where the smallest score (1) was attributed to all white animals and the largest (5) attributed to animals with black shades. The final score for each animal was decided based on majority voting. A total of 379 animals were evaluated, from which 131 were chosen and included in this study based on extreme scores. These animals were genotyped with the Illumina BovineHD Bead Chip assay, which included 777, 000 single nucleotide polymorphism (SNP) markers. After a standard genotype quality control, a genome-wide association analysis (GWAS) was conducted in order to map loci linked to the white phenotype. Considering a Bonferroni-corrected significance level of alpha = 1.15e-07, we detected highly significant associations (p = 7.636928e-127) mapping to chromosome 13, in the vicinity of the Agouti signaling protein gene (ASIP). Given that ASIP is a well-known color dilution gene, and that the expression pattern of ASIP in the coats of model organisms follows closely the distribution of grey shades found in Nellore cattle, the gene is a promising functional candidate. We are now sequencing the whole genomes of 8 Nellore bulls (4 black and 4 white) for scrutiny of the identified locus in the search for the causal mutation.

Funding: Beatriz Batista Trigo is supported by CAPES (Coordination for the Improvement of Higher Education Personnel) scholarship and Marco Milanesi was supported by grant 2016/05787-7, São Paulo Research Foundation (FAPESP)

’’’’’

Precise Identification and Genome Recovery of Viral Pathogens Through a Non-Specific Target Virome Clinical Test

Deyvid Amgarten, Fernanda Malta, Murilo Castro Cervato, Nair Hideko Muto, Pedro Sebe, João Renato Rebello Pinho

Laboratório de Técnicas Especiais, Hospital Israelita Albert Einstein

Abstract

Massive parallel sequencing techniques radically changed the diagnostic workflow by providing a quick and powerful tool for clinical diagnosis and precision medicine. The basis of rare diseases as well as clinically relevant mutations underlying some types of cancers have been precisely diagnosed and characterized using NGS techniques. In a similar way, etiological causes of infectious diseases, such as encephalitis, arboviroses and hepatitis have been precisely identified. In this work, we report the validation process of a new metagenomic protocol for unbiased identification of viral pathogens in clinical samples. The test is based on a two-step cDNA random amplification followed by Illumina sequencing. Data generated is used as input to a custom bioinformatic pipeline especially developed for delivering fast, accurate and clinician-friendly diagnosis. The validation process was performed according to College of American Pathologists (CAP) guidelines, evaluating accuracy, intra and inter-assay reproducibility, and limit of detection. Preliminary results show accuracy rates of 100% compared with standard PCR tests, total agreement among different assays and 104 virions/mL as limit of detection (similar to PCR limits). Moreover, we were able to recover high quality data to characterize almost complete or whole genomes from different viral pathogens with mean depth of coverage ranging from 100x to 8000x. Recovered genomes were also used to identify viral genotype and drug resistance variants, as well as to perform phylogenetics analysis. Altogether, these results show a multi-functional test to potentially replace several gold-standard molecular diagnosis techniques such as qualitative PCR and Sanger sequencing method in a unique protocol. We emphasize that the virome technique developed in this work is not limited to previous knowledge of the pathogen, allowing early detection of outbreaks due to novel pathogens and playing a key role in public health surveillance.

Funding:

,

A container-based pipeline for bacterial genome assembly and annotation

Felipe Marques de Almeida, Georgios Joannis Pappas Junior

Universidade de Brasília

Abstract

Advances in DNA sequencing technologies are reshaping bacterial genomics studies enabling chromosome level assemblies, at a fraction of cost and time, paving the way to population level genomic surveys. At the present, the computational analysis of sequencing data is the main hindrance to the field and withholds its move into mainstream clinical settings. To overcome this barrier we developed a complete container-based pipeline for bacterial genomics analysis, meaning that given raw sequencing data from multiple platforms (Illumina, Pacbio and Oxford Nanopore), it performs genome assembly and annotation, enabling identification and visualization of antibiotic resistance genes, virulence factors, prophages and integrative elements. In general terms the annotation phase of the pipeline can be executed in a few hours in a laptop. The assembly module, despite requiring a large amount of memory (>64 Gb RAM), can be executed in a day. The pipeline is designed to be modular, taking into account different analytical scenarios readily configured by the user. We also leverage the use of operating system virtualization meaning that there is no need for user installation of required pipeline components. When ready, all modules will be made available through GitHub. In conclusion, this pipeline offers a seamless exposition of computational tools to bridge the gap toward routine bacterial genomics.

Funding: Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e Universidade de Brasília (UnB)

,,,,,,,,,

Identification of non-homologous isofunctional and species- specific enzymes in Mycobacterium abscessus as possible therapeutic targets

Fernanda Cristina Medeiros de Oliveira, Philip N Suffys, Solange Alves Vinhas, Moisés Palaci, Pedro Henrique Campanini Cândido, Elizabeth Andrade Marques, Tania Folescu, Rafael Silva Duarte, Marcos Paulo Catanho de Souza, Ana Carolina Ramos Guimarães

Fundação Oswaldo Cruz, Instituto Oswaldo Cruz, Laboratório de Genética Molecular de Microrganismos, Rio de Janeiro, Brazil

Abstract

The genus Mycobacterium is composed of several pathogenic species. Many organisms belonging to this group are responsible for diseases, such as tuberculosis, leprosy and other serious infections. Mycobacterium abscessus (MABSC) is responsible for severe infections that are becoming common all over the world. Outbreaks caused by MABSC have already been reported in several countries, including Brazil. Treatment of infections caused by MABSC remains ineffective due to antibiotic resistance, justifying the search for potential new therapeutic targets. The identification of non-homologous isofunctional (analogous) and species-specific (taxonomically restricted) enzymes in these parasites compared to the human host might be a relevant approach to disclose potential new therapeutic targets, as we already demonstrated in previous works. Although catalysing the same reaction, non-homologous isofunctional enzymes have distinct evolutionary origin as well as significantly distinct fold topologies and three-dimensional structures. On the other hand, species-specific enzymes are those found only in one organism in relation to others, therefore comprising the most evident targets. In this work, eleven M. abscessus genomes were annotated with three distinct tools to identify and characterize encoded proteins with predicted enzymatic activity: RAST, Argot2.5, and BLASTKOALA. A computational pipeline developed by our group (AnEnPi) was used to predict genes encoding putative non-homologous isofunctional enzymes in enzymatic activities shared between those mycobacteria and the human host, as well as species-specific enzymes encoded in M. abscessus genomes. We identified events of evolutionary convergence in 11 enzymatic activities comprising 11 metabolic pathways, shared by M. abscessus and H. sapiens, involving 153 enzymes in total. Five enzymatic activities (EC 5.3.3.2 - terpenoid biosynthesis, EC 2.3.1.1 - arginine biosynthesis, EC 2.7.7.4 - purine metabolism, EC 4.1.1.19 - arginine and proline metabolism, and EC 4.2.2.1 - nitrogen metabolism) are candidates to be further investigated as new therapeutic targets for drug development against M. abscessus. On the other hand, 1, 592 species-specific enzymes (compared to the human host), shared among the eleven isolates of M. abscessus, were also identified in this process, belonging to 105 distinct enzymatic activities acting on 96 metabolic pathways of M. abscessus. The predicted non-homologous isofunctional instances were confirmed based on structural folding and protein signatures assigned to the implicated enzymes. Mapping these non-homologous isofunctional and species-specific enzymes in M. abscessus metabolism revealed potential new therapeutic targets to control infections caused by MABSC.

,,,

Impacts of retroelements in tumorigenesis

Fernanda Orpinelli, José Leonel Lemos Buzzo, Thiago Luiz Araujo Miller, Pedro A F Galante

USP

Abstract

Colorectal cancer (CRC) is the third most common cancer in the world, with nearly 1 million new cases annually diagnosed. Of these, about 50% of the patients evolve to death mainly due to the development of metastatic (secondary) tumors originated from primary colorectal tumors. Thus, studying genetic variations in secondary tumors is central to better understand tumor progression and to provide treatments that are more effective for CRC-affected patients. Here, we are using whole genome sequencing (WGS) data - matched normal tissue, primary and secondary tumor tissues from 5 CRC-affected patients - and bioinformatics methodologies to study genomic variation present CRC. Among genomic variations, we are developing methods to identify and evaluate the functional role from those caused by LINE-1, retrocopies of protein-coding genes and Alu elements, these latter two unexplored genomic variations with great potential to be functional. Three of five patients (60%) were stage IV when diagnosed. When comparing early-stage patients with late-stage in the progression of CRC, we found that the number of (likely) novels LINE-1 insertions are (10x) higher in the late stages of the disease. However, these numbers are not the same while looking for retrocopies of coding genes. Patients that were stage IV had on average 40 retrotransposition events, while one patient stage II had 95 retrotransposition events from coding genes and the stage III patient had only three. Thereby, our work aims to investigate two types of variations poorly studied, mostly in secondary/metastatic tumors, and contribute to an understanding of the frequency and importance of these variations in tumorigenesis and metastasis of a very relevant type of cancer, colorectal cancer.

Funding: FAPESP

,,,,

Original human genome might have had 25% to 35% methylated C in CpG

Fernanda Stussi, Carlos Alberto Xavier Gonçalves, Lissur Azevedo Orsine, Tetsu Sakamoto, J. Miguel Ortega

UFMG

Abstract

Some authors suggested an effect of neighbor bases on the probability of SNPs occurrence. We built a graphical online database SNP LocAL Neighborhood and computationally over deaminate every CpG to TpG, supposing that C was methylated and increased the bias or artificially aminate fractions of TpG to simulate reversion to CpG with methylated C. Aiming to investigate comprehensively this event, we built an online database to show the pattern of bases in SNP neighborhood, available at: http://bioinfo.icb.ufmg.br/snplane/. SNP LANE comprises SNPs in Mus musculus, Homo Sapiens, Bos taurus, Rattus Norvegicus and Sus scrofa, localized in intron, CDS, 5'UTR or 3'UTR and classified by substitution type: K, M, R, Y, W or S. For each SNP class, nucleotide frequencies were calculated for the first five positions upstream and downstream surrounding the SNP. Expected baseline nucleotide frequencies for positions neighboring the SNP were estimated by randomly choosing positions in the genome and retrieving nucleotides flanking it. Two graphics are presented for each of 1200 distinct situation. In the majority of cases baseline frequency was not significantly different from observed data, indicating that the observed neighboring effect was not an influence on the mutation, but rather if T or A are more frequent downstream of C, it would seem C might be influencing the transition T/A but baseline frequency indicates that this is just an effect of non-randomness of the genome. When we deaminated all remaining C in CpG, was a small increase in bias. Simulating different percentages of amination of "CpA" and "TpG" back to CpG dinucleotides was noteworthy that bias is completely erased with 25% to 35% of amination. We do not see the neighboring nucleotide effect on these conditions. R and Y substitutions did not respond to amination, probably because amination already causes R and Y. It is suggested that dinucleotide composition produces the previously reported neighborhood bias on SNP probability. Most of this effect might be explained by deamination of C in CpG and we suggest that originally human genome would have 25% to 35% of the present CpA and TpG in the form of CpG.

Funding: Fapemig, Rede Biologia Sistêmica do Câncer

,,,,

Assessment of intratumoral genetic heterogeneity scores (ITGH) and its association with clinical parameters across several cancer types

Filipe Ferreira dos Santos, Cibele Masotti, Isac de Castro, Anamaria A. Camargo, Pedro A F Galante

Molecular Oncology Center, Hospital Sírio-Libanês, São Paulo, Brazil

Abstract

In the age of precision medicine, the use of molecular data has become increasingly common in clinical oncology routine, especially with the advent of Cancer Gene Panels (CGPs). Given the relevant influence of intratumoral genetic heterogeneity (ITGH) on the prognosis and treatment of patients, its better understanding is vital. Recent advances in sequencing technologies and computational algorithms allowed the development of tools that estimate ITGH based on mutant allele frequencies (MAFs). However, several practical and methodological limitations make it difficult to get into the routine of clinical oncology. MATH score is a method capable of estimating ITGH from single biopsies taking into account parameters that are known to disrupt MAF estimates such as sample purity. This study aimed to measure the ITGH of all 33 cancer types from exome (WXS) data of the TCGA project and the MSK-IMPACT 410 cancer gene panel with MATH score in order to evaluate its association with several clinical parameters. Univariate survival analysis was done with stratified 5-Fold cross-validation strategy and ROC curves. For multivariate analysis, a modified cox regression model was used, joining the Monte-Carlo cross-validation strategy and the Bootstrap resampling method. ITGH varied significantly among patients and cancer types, where generally more aggressive tumors have higher levels of ITGH such as OV. In addition, MATH is a good prognostic marker of OS for UCEC, UCS and LGG and PFI for UCEC and LGG with WXS data. Similarly, significant results were obtained for OS in LUSC and for PFI in LUAD from CGP data. Additionally, higher MATH was associated with high TNM (stages III and IV) staging for UCEC (clinical), COAD, SKCM, KIRP, and ESCA (pathological) with WXS data and for colorectal cancer with CGP data. In addition, higher MATH was also related to the presence of metastasis in UCEC, ESCA, COAD, KIRP, and KIRC with WXS data. With MSK-IMPACT 410, higher ITGH was associated not only with presence, but also with a higher risk of metastasis in patients with colorectal cancer. On the other hand, MATH was not significantly different between responders and non-responders to immunotherapy in both the WXS and CGP data. Therefore, MATH presents itself as a promising tool for oncologists due to its simple use and easily interpreted results, in addition to the aforementioned associations with survival, staging and metastasis. In conclusion, MATH may have several applications in the near future regarding patient prognosis and therapeutic decision making.

Funding: Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001, Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) - processo n^{o} 2017/17974-9

,,,,,,

Comparative genomics of Acinetobacter baumannii strains

Diego Lucas Neres Rodrigues, Raquel Enma Hurtado Castillo, Daniella Camargo Costa, anne cybelle pinto gomide, Vasco A de C Azevedo, Francielly Rodrigues da Costa, Flavia Figueira Aburjaile

UFMG

Abstract

Acinetobacter baumannii is an important opportunistic pathogen causing meningitis, bacteremia, pneumonia, erysipelas and genitourinary tract infections. In recent years, this etiologic agent has acquired multiple mechanisms of resistance to a diverse range of antimicrobials. Its ability to survive in different environments combined with its resistance to drugs makes it extremely difficult to treat patients suffering from infections associated with this pathogen. In this context, this research aims to elucidate in silico analysis from the genome of different strains of A. baumannii to elucidate some mechanisms of resistance related to multiple drug efflux pumps, as well as, to search for genomic resistance islands among 122 A. baumannii strains. By means of the in silico comparison of fifteen different strains of A. baumannii, it was possible to observe a phylogenetic approximation of the A. baumannii individuals, with the presence of few polymorphic points in the conserved 16S rRNA gene, as well as the presence and metabolic analysis of 14 proteins related to efflux pumps in each of the selected strains, all of which were evaluated after automatic annotation and manual curation, where the results maintained the functionality of the proteins with reliability. In conclusion, there is great phylogenetic proximity between the A. baumannii strains studied. However, regarding comparative genomics and sequence annotation, there are differences between the products generated automatically and manually, showing possible points of polymorphism. In addition, the search for metabolic islands revealed the presence of resistance islands that may elucidate the prolonged survival of this bacterial species.

Funding: CAPES, CNPq e FAPEMIG

,,,

Characterization of Xop effectors in Xanthomonas citri subsp. malvacearum

Manuela Correia Dionísio, Juan Carlos Ariute, Ana Maria Benko-Iseppon, Flavia Figueira Aburjaile

UFPE

Abstract

Cotton (Gossypium spp.) is a fiber cash crop extensively sown throughout the world. At the Northeast of Brazil, for example, cotton yields are highly valuable for export. However, in the last years, the producers have been challenged by the maintenance costs of the plant. Once that diseases and pests significantly reduce cotton yield, and a phytosanitary control must be implemented. In plant pathogenic bacteria, type III effectors (T3Es) play a crucial role in pathogenicity, mainly against Gram-negative bacteria. Bacterial blight of cotton is incited by Xanthomonas citri subsp. malvacearum, a disease responsible for large losses of cultivars. Chemicals products used to treat the disease present low efficacy and pollute the environment. Therefore, targeting the group of effectors like Xops (Xanthomonas outer proteins) constitutes a significant approach for the generation of more resistant plants. In this context, this study aims to identify and characterize Xops genes in genomes of X. citri subsp. malvacearum. Therefore, we selected seven genomic sequences of X. citri subsp. malvacearum in GenBank (NCBI database). Afterward, an automatic annotation was carried out using RAST (Rapid Annotation using Subsystem Technology) tool followed by manual curation. Further, we intend to obtain the Xops involved in the pathogenicity of cotton. During the automatic annotation (RAST), we found some plasmids and transposable elements, which explains their high genetic variability. Also, we suggested that virulence genes are implicated in plant-pathogen interaction with effectors type III (T3Es). Thus, we identified effectors like xopC, xopJ, xopAG, AvrBs3, AvrXa10 and PthXol in X. citri pv. malvacearum in the 7 genomes and we related to their pathogenicity. In our studies, we show the presence of Xop effectors, as well as their dynamic organization in the genomes. Furthermore, the analysis of comparative genomes by Mauve software presented deletions, inversions and insertions, which suggest the genetic variability in these strains. Hence, the characterization of these proteins may subsequently be used, as targets for the control of phytosanitary problems in cotton the Brazilian Northeast.

Funding: CAPES, CNPq, FACEPE.

”

SEARCH AND CHARACTERIZATION OF NON-CODIFYING RNAs IN ISOLATES OF Bacillus thuringiensis (Bacillus cereus sensu lato) BY RNA SEQUENCING

Viviane Aparecida Gobetti, freddy Eddinson Ninaja Zegarra, Laurival Antônio Vilas Boas

Universidade Tecnológica Federal do Paraná

Abstract

Evidence shows a close similarity in the genomes of Bacillus thuringiensis, Bacillus cereus and Bacillus anthracis bacteria belonging to the Bacillus cereus Sensu Lato group, suggesting that they should be considered to be of the same species. Thus, this research work aims to identify and characterize non-coding RNAs in Bacillus thuringiensis (Bt) strain 407, assisting in the taxonomic understanding and understanding of virulence factors through the proposed methodology, which adds search for homologous sequences in Rfam database. The adopted methodology used the complete genome of the Bt 407 strain of B. thuringiensis along with the total sequencing of Bt enriched with small RNAs extracted in two growth phases of the bacteria (end of log growth phase and beginning of stationary phase respectively). Subsequently, genome sequences were extracted in highly expressed regions, where possibly there would be a greater likelihood of finding candidates for ncRNAs. The extracted sequences were used together with the cmsearch tool of the INFERNAL package, using the maximum inclusion limits of 0.01 e-value for candidate classification. At the end of the research, a total of 894 candidates were identified, with 355 significant candidates (according to inclusion limits) and 49 ncRNA families. The work allowed the search and identification of ncRNAs in the Bt 407 strain of B. thuringiensis. , demonstrating different composition profiles of ncRNAs, with variations of candidates and sequences in both phases of growth of the bacteria, contributing to a better understanding of the life cycle of this bacterium, as well as the ncRNAs responsible for certain characteristics presented by this species.

Funding:

,,,,,,,,,,

The Death Is Red: Analysis of the Predicted Secretome of Aspergillus welwitschiae, with Emphasis in Pathogenicity and Carbohydrate Metabolism

Gabriel Quintanilha Peixoto, Daniel Silva Araújo, Rodrigo Bentes Kato, Paula Luize Camargos Fonseca, Luiz Marcelo Ribeiro Tomé, Fábio Malcher Miranda, Rommel Thiago Jucá Ramos, Bertram Brenig, Vasco A de C Azevedo, Fernanda Badotti, Eric Roberto Guimarães Rocha Aguiar, Aristóteles Góes Neto

University Göttingen

Abstract

In 2018 Aspergillus welwitschiae was described as the causing agent of the bole rot disease of sisal (Agave sisalana) rather than Aspergillus niger as previously thought. A. welwitschiae is a cryptic species of the A. niger/welwitschiae clade. Since then, we have sought to understand the mechanisms of pathogenicity of this fungus by sequencing the genome of two strains isolated from infected stem tissues of A. sisalana, and subsequent analysis of the gene content of these fungi. The genomes, sequenced in Illumina HiSeq 2500 platform, were assembled de novo and annotated. Obtained genome size in CCMB663 was 34.8 Mbp containing 12, 549 protein-coding genes, while CCMB674 genome is 32.2 Mbp long, containing 11, 809 protein-coding genes. The resulting gene prediction models were used for analyzing secretomes, in which different software was applied as filters to select the putative secreted proteins that were subsequently associated with Gene Ontology (GO) terms. Our results show that most of the associated terms describe functions associated with the degradation of proteins, lipids, and especially carbohydrates and that the relative abundance of these terms is similar between different isolates. From this set of proteins predicted to be secreted, possible effector proteins were identified, which are involved in the invasion and colonization of the infected plant. Most of the identified putative effector proteins have no similarity with known effectors, and in those in which was possible to identify conserved domains, essential and very important genes associated to virulence in plants are included, including genes that act to silence the immune response or to stimulate it, causing cell death in plant tissues. Secretory and cytosolic plant cell wall degrading enzymes were also analyzed, describing a great diversity of genes encoding for carbohydrate-active enzymes, some of which specialized in the degradation of carbohydrates present in large amounts in the tissues of sisal stem. Altogether, our results improved the understanding of the Aspergillus welwitschiae x Agave sisalana pathosystem, and also identified new effectors of interest in the fungus-plant interaction.

Funding:

”

A study of genetic diversity of Escherichia coli BH100 through structural and comparative genomics

Gustavo Santos de Oliveira, Andreia Maria Amaral Nascimento, Edmar Chartone de Souza

Abstract

Genetic variability can be seen as the driving force to evolution. In prokaryotes, many mechanisms emerged such as microorganisms could enhance variability, allowing their spread throughout different ecological niches. In a clinical context, major attention is given to the capacity of some microorganisms in colonize human tissues and developing diseases. From many different bacterial infections currently known, the urinary tract infection (UTI), which is caused mainly by Escherichia coli, can be highlighted as a major concern for public health. The comprehension of mechanisms associated with the emergency of variability and pathogenicity only was possible thanks to the advancements of Molecular Biology and more recently to Bioinformatics. In the present work, we aimed to use and develop bioinformatic tools in order to assembly, annotate and fully characterize the complete genome of E. coli BH100, which was isolated in 1973 in Belo Horizonte, Brazil, from urine of a patient suffering from UTI. This strain is resistant to ampicillin, tetracycline, kanamycin, chloramphenicol, inorganic mercury and in some cases to streptomycin. This multiresistance is due the presence of a mobilizable plasmid carrying a beta-lactamase gene (bla) and a self-transmissible R plasmid carrying the genetic resistance marks of all other elements. Thanks to the use of Next Generation Sequencing (NGS) we were able to assemble the genome of this strain as well as other variants built from curation of the original plasmids, such as we could interrogate the effects of these elements to the emergence of variability. The complete genome of E. coli BH100 resulted in a chromosome of $\tilde{5}$.2 Mb, the smaller mobilizable plasmid of $\tilde{1}$5 kb, and a self-transmissible plasmid of $\tilde{1}$07 kb. This strain has shown a considerable number of insertion sequences from the family IS3 spread differently along the chromosome of BH100 and its variants. The comparative analysis indicate that this strain might be an authentic uropathogenic E. coli (UPEC) causing pyelonephritis. The functional annotation confirmed the presence of all resistance marks in transposons (Tn). Special attention was given to the presence of Tn21, identical to the one found in plasmid NR1, and to a potential new transposon carrying a gene for kanamycin resistance flanked by IS5 elements. In the end, we propose a new mechanism capable to explain the emergence of unstable and diversified resistance to streptomycin within the population of E. coli BH100. In this model, we suggest streptomycin resistance shows up due an increase in the copy number of the gene aadA1 present in Tn21

Funding:

””

Comparative genomics in the search for Antifreeze Proteins in Metschnikowia australis

Heron Hilário, Thiago Mafra Batista, Carlos Augusto Rosa, Luiz Henrique Rosa, Glória Regina Franco

Universidade Federal do Sul da Bahia

Abstract

Metschnikowia australis is a marine yeast confined to the Antarctic costal region. Conversely, its most related species, Metschnikowia bicuspidata, is worldwide distributed, but rarely retrieved from Antarctica's seas. Besides global adaptations to life at low temperatures, many psychrophilic organisms, from bacteria to vertebrates, have independently evolved protein coding genes that directly interfere with the ice formation process. These genes are collectively called Antifreeze Proteins (AFPs), consisting of a diverse class that have inumerous biotechnological applications, from frozen food industry to organ preservation for transplants. As polyphiletic, AFP genes are harder to find with traditional sequence similarity search tools in less characterized taxa, as Metschnikowia. One M. australis specimen was isolated by the MycoAntar project and, due to its ability to survive freezing, it was selected for further studies by our group. We have sequenced and assembled the M. australis genome, for further genomic investigation. As M. bicuspidata genome was already available, we devised an strategy to predict Open Reading Frames (ORFs) on both genomes and select those exclusive to M. australis, which could potentially code new AFPs. The 249 ORFs exclusive to australis were then submmited to three machine learning based AFPs classifiers, in order to select the most promissing candidates for in vitro expression analysis. Primers were designed for 17 selected candidates, and these are now being used to probe both yeasts transcriptome after freezing stress. The candidate ORFs that are proven to be expressed will be further characterized. Genomic deletion and heterologous expression techniques will be used to confirm their relation to freezeing survival.

Funding: Capes

”

Using Drosophila melanogaster Y chromosome heterochromatic sequences as a model to construct complete oligopaints

Isabela Pimentel de Almeida, Maria Dulcetti Vibranovski, Antonio Bernardo de Carvalho

Universidade Federal do Rio de Janeiro

Abstract

Researches that focuses on understanding the sequence and organization of heterochromatin allows essential functions for the organism to be better understood. One of the main obstacles in studies with the Y chromosome is related to the heterochromatic state of this structure, which in Drosophila melanogaster is formed by approximately 41 Mb of highly repetitive sequences. With applications in cytogenetic studies, the protocol for constructing oligopaints probes does not usually include such sequences. Considering that there are only 1381 probes for the D. melanogaster Y chromosome, whereas for its other chromosomes this value is at least ten times greater, oligopaints for this structure do not allow analyzes as deep as for others. Using D. melanogaster Y chromosome as a model, we aim to identify unique repetitive sequences of it through the YGS (Y chromosome Genome Scan) method, increasing the number of known probes for this structure and allowing to construct its complete oligopaint. Thus, it is possible to compile the methodologies involved in the development of a new technique that will allow to construct oligopaints of the Y chromosome of any species of interest. The analyses of the results and the own efficiency and quality of the obtained oligopaints will be given through the direct comparison between these and those generated from the 1381 probes currently known for the Y chromosome.

Funding: CAPES e FAPESP (Jovem Pesquisador 2015/20844-4)

”

Warfarin dosing prediction in Brazilian patients using Algorithms based on Regression and Neural Network Models

Jennifer Eliana Montoya Neyra, Paulo Caleb J. L. Santos, Júlia Maria Pavan Soler

Laboratory of Genetics and Molecular Cardiology, Faculdade de Medicina FMUSP, Heart Institute (InCor), University of São Paulo.

Abstract

To analyze the performance of warfarin dosing prediction through Multiple Linear Regression (MLR) and MultiLayer Perceptron (MLP) algorithms in Brazilian patients from the Heart Institute (InCor-USP), were used demographic, genetic and clinical informations of 749 individuals with a maintenance doses in a stable state of warfarin. In addition, International Normalized Ratio (INR) values, between 2 and 3, were used to monitor how well the blood-thinning medication is working to prevent blood clots. The dataset was partitioned in 599 individuals to the training group and 150 individuals to the test group. From the available characteristics, 16 variables corresponding to the IWPC algorithm were evaluated and the result compared with the set of all 74 variables analyzed under the MLR and MLP algorithms. The mean absolute error (MAE) was used to assess the accuracy of the models, which is a metric widely used in warfarin prediction studies. The results show for 74 variables that the MLP had a better performance (MAE = 7.87 mg/week, SD = 10.06) compared to the MLR using the same variables (MAE = 8.49 mg/week, SD = 11.13), and also showed better results that using the variables proposed by the IWPC for both the MLR (MAE = 7.99 mg/week, SD = 10.86) and MLP (MAE = 8.44 mg/week, SD = 11.06) models. The performance of the MLP and MLR models tested in this study showed the well-known tendency of the MLP model to obtain better results when are analyzed a greater number of characteristics. This allows us to consider this type of neural network as a good candidate for the prediction of warfarin maintenance doses, taking into account that at present about of 600 variables have been related to anticoagulant therapy.

Funding: CAPES

,,,,,,

Genome-wide Association Studies Reveal Candidate Genes Important for the Interaction of Bacillus pumilus with Arabidopsis thaliana

Marina Soneghett Cotta, Fernanda do Amaral, Leonardo Magalhães Cruz, Fabio de Oliveira Pedrosa, Emanuel Maltempi de Souza, Tadashi Yokoyama, Gary Stacey

University of Missouri

Abstract

The plant growth promoting bacterium (PGPB) Bacillus pumilus is a nitrogen fixing and a gibberellin producer that increases the nitrogen content and shoot length and surface in plants. In addition, this PGPB has the capability of improving plant growth under drought and saline conditions. The strain Bacillus pumilus TUAT-1 can increase rice's roots and biomass and the content of nitrogen and chlorophyll. In this study, through genome-wide association study (GWAS), we evaluated the interaction between TUAT-1 and Arabidopsis thaliana. In order to do that, 288 A. thaliana accessions were screened for root architecture traits: main root length (MRL), number of lateral roots (NLR), branched zone (BZ), total root length (TRL) and lateral root length (LRL). GWAS accelerate mixed model analysis was performed for the 5 traits within 288 ecotypes. Several ecotypes responded significantly to TUAT-1 inoculation: MRL (52.7%), NLR (14.2%), BZ (8.3%), TRL (21.2%) and LRL (19.1%). Many inoculated ecotypes were significantly affected in more than one trait. However, only one ecotype showed a significant difference between control and inoculated plants for all the traits. While some of the ecotypes either did not respond or responded positively for growth, a few ecotypes showed inhibition of root growth upon inoculation. Significant single nucleotide polymorphisms (SNPs) were detected in all the traits evaluated. 34% of the significant SNPs were associated with more than one trait and, 1 SNP was associated with 4 different traits. Causative SNPs were selected according to missense or nonsense alterations and produced a list of candidate genes related to hormone production, defense response, root development, autophagy, and fatty acid metabolism. Candidate genes were very likely associated with the interaction between TUAT-1 and Arabidopsis and the plant growth promotion. In this study, we validated previous reported Bacillus spp. and its plant interaction and growth promotion genes and highlight the potential genes involved in these mechanisms. The results of this work showed that some of the root architecture characteristics are genetic separable traits associated with the plant growth. We suggest that plant-bacteria interaction and the plant growth promotion are quantitative and multigene traits. This knowledge expands our understanding of the functional mechanisms driving the plant growth promotion by PGPB.

Funding: Supported by INCT-FBN, CNPq, Fundação Araucária and CAPES

”

Targeting audience and tailoring courses using the ISCB Competency Framework: An application survey from RSG-Brazil Educational Committee

Maira Rodrigues de Camargo Neves, Raquel Riyuzo de Almeida Franco, Nilson Coimbra

Universidade Federal de Minas Gerais

Abstract

Reactivated in 2015, the ISCB Regional Student Group Brazil (RSG-Brazil) is a vibrant student network with a mission to aid the training of the next-generation of bioinformaticians and computational biologists among Brazil. This student group has the support of the Student Council from the International Society of Computational Biology (ISCB-SC), which is dedicated to advancing the scientific understanding of living systems through computation. In view of its mission, the Educational Committee of RSG-Brazil released an ongoing survey, released in April of 2018, to identify complementary needs on the ongoing formation in the field of bioinformatics and computational biology. In this work, we re-devised the ISCB competency framework, an international effort that identified 16 core competencies required by professional working in the fields related to computational biology, to only 5 categories in order to increase the community engagement on the form, as well as reduce the time demands for respondents. Respondents were asked to declare interest in learning more about one or more of the 9 pre-defined topics of Bioinformatics shown in the form, or to make new suggestions. Also, target audience was asked to identify themselves within one of the three profiles of Bioinformatics audience and their roles or subcategories, as defined by the ISCB Competency Framework (Bioinformatics user, Bioinformatics scientist and Bioinformatics developer, and their subcategories). The survey was released in May of 2018 and has so far collected 216 responses, from 21 (out of 26) Brazil states, with over 85% engagement from undergrad or graduate students. We have found that most of the respondents declared as Bioinformatics users, working in academia. We also have identified particularities in the interest in topics by country region and by Bioinformatics profile. With this effort of the Educational Committee, we expect to establish partnership between the Regional Student Group Brazil and universities/companies to delivery tailored courses to the student community of computational biology in Brazil.

Funding:

”

Antibiotic resistance genes in the gut microbiome of worldwide populations

Liliane Conteville, Gregorio Manuel Iraola Bentancor, Ana Carolina Paulo Vicente

Oswaldo Cruz Institute

Abstract

The human lifestyle and the environment have a direct impact not only on the taxonomic and functional profiles of the human gut microbiome but also on its collection of antibiotic resistance genes (ARGs), the resistome. ARGs have been identified even in the microbiomes of human populations that were never exposed to commercial antibiotics. This is correlated with the fact that microbial resistance has always been naturally occurring in the environment. To explore the resistome profiles of distinct populations worldwide, we analyzed 1072 human gut microbiomes from 21 human populations with different diets, lifestyles, and genetic backgrounds. This study is original considering the metagenomes dataset analyzed and the approach performed. The programs ARIBA and ABRICATE were used to screen for ARGs in the metagenomes with paired-end and single-end reads, respectively. The ARGs identified were grouped and sequence redundancy was removed using CD-HIT, which generated an ARGs catalogue with 328 sequences. Each metagenome was mapped against this catalogue with BBMAP. DESEq was used to normalize the counts of the reads mapped. PcoA and cluster analysis showed a discernible separation among westernized and non-westernized groups. Genes that confer resistance to tetracycline are the most prevalent genes in most of the westernized and non-westernized groups, but the westernized groups show a higher abundance of these genes. Considering the 21 groups, it was possible to identify 182 significantly discriminative features among them. Interestingly, relevant ARGs commonly found in clinical isolates were found in remote or semi-isolated groups, as TEM, OXA, Cfx, aadS. The characterization of the resistome of worldwide populations and the understanding of the routes of ARGs spread are particularly relevant to global public health.

Funding: CAPES, CNPQ, IOC, PASTEUR

,

Scientific Dissemination in Bioinformatics

Luana Luiza Bastos, Raquel Melo Minardi

Universidade Federal de Minas Gerais - UFMG

Abstract

Computational science has greatly contributed to the generation and organization of biological data as well as the development of more sophisticated techniques for solving biological problems. From gene sequencing to molecular biology the machine work has facilitated the discovery of revolutionary informations. As knowledge is accumulated and the need for understanding of life processes has increased, biology keeps its position as a central science with greater reach and impact on society. Bioinformatics has gained greater relevance since it is the interdisciplinary field of knowledge linking biology with computation. This new science is responsible for the increasingly contribution to the development of algorithms, that are unfortunately still applied and studied by a small number of interested people. Life science (and related) researchers have high demand for the development and use of these algorithms as they effectively contribute to the quality of the data produced. However, often the algorithms used are poorly understood and are applied without the proper domain of their operation and parameters. In addition, much has been said about the importance of scientific dissemination as the general population is far out removed from the researches carried out at the university. In this context, the aim of this project is to develop didactic content to make bioinformatics algorithms simpler to understand and make the processes involved in the development more transparent to the users of a respective software. Besides, we would like to foster the creation of communities for research discussion between researcher and society on topics related to bioinformatics in social media, as well as to produce teaching content and scientific dissemination about bioinformatics. For this, a channel on the Youtube platform were created with the proposal to present videos related with bioinformatics informations. Also, it has been created a web site in which are deposited basic texts and references used for the production of videos and where accounts on social medias platforms, such as Instagram and Facebook, are divulgated in which content about bioinformatics and scientific dissemination are deposited periodically. A research on bioinformatics was conducted with undergraduate and graduate students of the Federal University of Minas Gerais through an online based form, about the most used and impacting algorithms in different areas of bioinformatics. In addition, this work aimed to understand the demands presented by students and themes that would be of interest to video production. For the selected themes, related content is produced for social media, including videos and text production. Currently the channel on Youtube called OnlineBioinfo has about 60 videos posted and has over 850 subscribers with approximately 21, 621 views. In the future it is expected to increase the number of people reached by the content created, besides to improve its quality by developing guidelines that facilitate learning and that are of interest to the community.

Funding:

,,,,,,,,,,,,,,

Taxonomy and comparative genomics of a Corynebacterium ulcerans strain isolated from Pig, previously identified as C. pseudotuberculosis

Janaína Canário Cerqueira, Rodrigo Profeta Silveira Santos, Alessandra Lima da Silva, Raquel Enma Hurtado Castillo, Marcelle Oliveira Almeida, Thiago de Jesus Sousa, Diego Lucas Neres Rodrigues, Juan Luis Valdez Baez, Francielly Rodrigues da Costa, anne cybelle pinto gomide, Henrique Figueiredo, Alice Rebecca Wattam, Artur Silva, Vasco A de C Azevedo, Marcus Vinicius Canário Viana

University of Virginia

Abstract

The bacterial strain PO100/5 was isolated from a skin abscess of a pig (Sus scrofa domesticus) in the Alentejo region of southern Portugal. It was identified as Corynebacterium pseudotuberculosis using biochemical tests (Api Coryne© kit), multiplex PCR and Pulsed Field Gel Electrophoresis. After genome sequencing and in silico analyses, the strain was re-identified as C. ulcerans and deposited in Genbank. This species can harbor the Phospholipase D (PLD) toxin and diphtheria toxin (DT), has variety of mammalian hosts as reservoirs, and emerged in the last 30 years as the main cause of diphtheria in humans. To better understand the taxonomy of C. ulcerans and improve the identification methods, we compared strain PO100/5 to other Corynebacterium genomes publicly available. The taxonomic identification of this strain was refined using 16S gene sequence identity, phylogenetic analysis of the genes 16S and rpoB, phylogenomic analysis using the nucleotide sequence of 1, 000 shared genes, and the Average Nucleotide Identity (ANI). The identity of the 16S gene was calculated in relation to the type strains C. ulcerans NCTC7910 (99.673%), C. pseudotuberculosis ATC19410 (99.476%) and C. dipthteriae NCTC11397 (97.709%). Those values were above the 97% cutoff used to identify the same species. Two strains had 100% identity with PO100/5: C. ulcerans W25 (wild boar, Germany) and KL1196 (deer, Germany). All phylogenies (16S, rpoB and phylogenomics) showed a clade composed of C. ulcerans strains PO100/5, W25 and KL1196 separated from the other C. ulcerans strains and from the closest species C. pseudotuberculosis and C. diphtheriae. Additionally, C. ulcerans strains were separated in three clades. The ANI analysis showed that within strains PO100/5, W25 and KL1199 the pairwise values varied from 99.59 to 100%, and from 90.13 to 90.34% when those were compared to other C. ulcerans strains. Within the other C. ulcerans strains, those values varied from 95.32 to 99.94%. According to ANI method, genomes with an identity below 95% are considered different species. The genome plasticity was accessed by prophage and genomic island prediction. Four prophages were predicted in the genome of strain PO100/5, one of them harboring the DT gene (tox). Eight and sixteen genomic islands were predicted by comparing PO100/5 to the type strains C. ulcerans NCTC7910 and C. pseudotuberculosis ACTT19410, respectively. The genomic circular map showed three regions shared by strains PO100/5, W25 and KL1199 that are absent in other C. ulcerans and the closest species C. pseudotuberculosis and C. diphtheriae. These regions could be used as molecular markers for identification by multiplex PCR. The results suggest that strains PO100/5, W25 and

,

NLR genes in aquatic mammals and where to find them

Maria Luiza Andreani, Mariana Freitas Nery

Unicamp

Abstract

The immune system relies on receptors to achieve a proper immune response. Pattern Recognition Receptors (PRRs) are the main receptors in the innate immune response. PRRs allow the detection of conserved pathogen-associated molecular patterns (PAMPs) and damage-associated molecular patterns (DAMPs). As a survival strategy, pathogens may develop mechanisms to avoid detection by PRRs and, therefore, there is room for pathogen mediated selection. Changes on ecological environment may lead to new challenges to immunity system, as composition of pathogens may vary. Accordingly, here we analyzed an innate immune family of receptors in our own sequenced genomes of non-model species of aquatic mammals, comparing recently diverged species of fluvial and marine mammals in cetaceans and sirenians. We focused on the NLR (Nod-Like Receptors) family of immune genes formed by a NACTH domain, a C-Terminal LRR region and a CARD or PYD domain in the N-Terminal region. They are intracellular receptors that can detect signs of viruses, bacteria and cellular damage. From this family we characterized the NOD1, NOD2, NLRC4 and NLRP3 genes in Sotalia guianensis, Sotalia fluviatilis, Trichechus manatus and Trichechus inunguis. We blasted the corresponding Tursiops truncatus genes on our sequenced genomes and investigated 10, 000 bp upstream and downstream searching for pseudogenes or duplicates as this family of immune receptors is known to vary on copy number along lineages. Moreover, NLR receptors NLRP2, NLRP4, NLRP5, NLRP7-9 and NLRP11-13 are located in the same chromosome in humans, therefore we identified the flanking genes of the most upstream and downstream receptors in ungulates species and blasted them in the sequenced genomes, extracting the region in between. Further, we compared our findings to genes present on terrestrial mammals. Characterization of innate immune genes on non-model species may lead to a better understanding of molecular evolutionary dynamics of host-pathogen interactions and how they evolve in the wild.

Funding: CAPES, Instituto de Biologia

”

Investigating the genomics of the evolution of sociality in Hymenoptera

Maycon Douglas de Oliveira, José Eustáquio dos Santos Júnior, Francisco Pereira Lobo

Universidade Federal de Minas Gerais

Abstract

Different forms of social behavior arose across metazoans, varying from simple aggregates to caste-differentiated insect colonies. Wasps, ants, and bees, insects of the Order Hymenoptera, have a wide variation in sociality, ranging from solitary individuals to colonies with hundreds of millions of individuals and the most complex form of social behavior, eusociality, defined by traits such as reproduction-based division of labor and cooperative brood care. One of the possible ways of annotating genes is by associating those with their biological functions, which are descriptive terms of the possible roles of that gene's products in biological systems. In this work, using the order of magnitude of individuals per colony (IPC) as a proxy for society complexity, we aimed at surveying high-quality Hymenoptera genomes for biological functions whose frequencies are significantly associated with the increase of colony size, searching for possible genomic components that may contribute to the emergence of this complex phenotype. After a thorough literature and database review, we selected 39 Hymenoptera species that have both high quality non-redundant proteomes (defined as the sets of the longest coding sequences per locus with an expected content of single-copy orthologs > 90%) and estimated IPC values. Each proteome was de novo annotated using InterProScan to predict protein domains using the Pfam database. Using this data, we searched for biological functions whose frequencies across genomes are significantly associated with IPC values. For significance, we took into account the multiple hypothesis scenarios for both Pearson's correlation and phylogeny-aware linear models, requiring corrected p-values < 0.1 for both tests. From the set of 4520 distinct biological functions, 11 were found to have a significant positive correlation. Some of them are proposed to play important roles in eusocial insects, such as GO:0016575 (histone deacetylation) and GO:0004407 (histone deacetylase activity). Epigenetic transcription control plays an important role in eusocial insects, regulating caste differentiation in Hymenoptera like Apis mellifera. Other biological functions and protein domains found to have correlations with higher degrees of sociality remain to be characterized in this context, and are interesting candidates to be subject of future research. We look forward to keep investigating those results and possibly shed new light on the genomics of eusociality in Hymenoptera.

Funding: PPG Genética - ICB - UFMG

””

Classification of Transposable Elements through Convolutional Neural Networks

Murilo Horacio Pereira da Cruz, Douglas Silva Domingues, Priscila T M Saito, Alexandre R Paschoal, Pedro Henrique Bugatti

São Paulo State University

Abstract

Transposable Elements (TEs) are the most represented sequences occurring in eukaryotes. They can change their location and generate multiple copies of themselves throughout genomes. This action can cause significant effects in organisms, such as the regulation of gene expression. There are several types of these elements which are classified into two classes, nine orders, and 29 superfamilies. Sequences are classified in a hierarchic way, in which classes are divided into orders and orders into superfamilies. The correct classification of these sequences is still a challenging quest. Due to the rapid increase in the number of sequenced genomes, the manual classification of these sequences is no longer feasible. Besides, automatic methods are mostly based on sequence alignment, a strategy that demands high computational costs. Therefore, novel strategies are required for this problem. To fill this gap, we present an automatic TE classification approach through a Convolutional Neural Network (CNN). Unlike traditional machine learning algorithms, that use handcrafted features to classify data, CNN can learn the best representation for the data and how to correctly classify it, given it is a representation learning algorithm. Few methods on the literature provide the classification of these sequences into the superfamily level. Superfamilies of the same order tend to share similar structures, i.e. the type of repeats, sequence length, and protein domains, thus providing a challenge to handcrafted features based methods. We evaluate the performance of CNNs on the classification of three datasets built using sequences from seven databases. We compared our results to TEclass, a method that uses Support Vector Machines to classify TEs into one class and three orders. Our approach obtained an accuracy of 92.3% on the classification of RepBase sequences from nine superfamilies and 94.6% on the classification of sequences from all seven databases into 10 superfamilies. We also obtained 98.1% on the classification of three RepBase orders and 94.3% on the classification of sequences from all databases from four different orders.

Funding: This work is supported by CAPES

,,

IN SILICO ANALYSIS OF THE CONSERVATION OF LEUCISM-RELATED GENES IN VERTEBRATES

Letícia Xavier Silva Cantão, Raquel Melo Minardi, Fabiana Alves

Universidade Federal de Minas Gerais - UFMG

Abstract

Leucism is an anomaly of the pigmentation of the skin of animals and manifests itself as the total or partial loss of the natural color of the species, and can affect parts of or the entire body of an individual. This change is caused by gene mutation, or by changes in expression in the some genes related to melanin synthesis. Related to this anomaly are the genes EDN3, EDNRB, KIT, MITF, PAX3 and SOX10, responsible for the migration and differentiation of melanocytes; this genes are highly conserved and indicate that they have some important function for the survival of organisms. The aim of the present study was to analyze in silico the conservation of Leucism-related genes in vertebrates and to evaluate the phylogenetic relationship of the same. The NCBI database was used to obtain vertebrate mRNA sequences. Then, global and local alignment was performed the significance of the data by E-value. Phylogenetic analysis was based on the construction of a Bayesian inference phylogenetic tree. Among Leucism-related genes, only EDN3 was significantly conserved.Possibly EDN3 is the main candidate gene for the leucism induction. It was found that the phylogeny of mammals selected (after alignments) for tree construction did not allow a well resolved relationship. While for the order Rodentia the relations corroborate with phylogenies already found in the literature. Birds and mammals were grouped into distinct groups.

Funding:

,,,,,

The role of host genetic variability in the development and establishment of human gut microbiome diversity

Ondina Fonseca de Jesus Palmeira, Larissa Matos, Michel Satya Naslavsky, Heloísa Bueno, Mayana Zatz, João Carlos Setubal

Institute of Mathematics and Statistics - University of São Paulo

Abstract

Studies have shown that human microbiome plays important roles in physiology, from food digestion to mental diseases. Since the gut microbiome composes the highest number of microbial cells outnumbering even our own cell counts, it is expected that the gut microbiome would affect a great deal of human biological functions. This makes the gut microbiome key to maintain homeostasis in the various biological levels where there are constant and active interactions between microbes, tissue, cells and molecules. The structure of the gut microbiota is shaped by many factors, including host genetics. Understanding how these factors determine the microbiome during the development and establishment of the gut microbiota at early stages of human life is crucial to infer biological and pathological microbiome composition. The purpose of this study is to apply tools of bioinformatics to process and analyse data samples of feces of triplet babies in order to verify host genetic associations with gut microbiome during the first 3 years of the babies lives. Infant feces are collected from the first week of life until 3 years of age. DNA extraction from the samples is performed followed by PCR (Polymerase Chain Reaction) targeting the specific bacterial 16S ribosomal RNA gene, region V3 - V4. Next-Generation sequencing is applied. Amplicons are then computationally processed by efficient algorithms to yield high quality reads. Later, these reads are assigned to groups of taxa based on differences on single nucleotide. Species specific classifiers and database are used for taxonomic assignment. Alpha and beta diversity are analysed with both qiime2 plugins and R packages. Most of our samples presented enough reads to identify all taxa. Phylogenetic diversity increased in samples on later time points presenting time as the dominant factor to determine alpha and beta diversity. Actinobacteria, Verrucomicrobia, Proteobacteria, Firmicutes and Bacteroidetes were the dominant phyla in all samples. At the species level, monozygotic twins presented more similar microbiome between them than between their dizygotic sibling. Our results showed that genetic factors could be detected at the species level and that time is crucial for diversity in the first months of life. The identification of the gut microbiome structure in both monozygotic and dizygotic twins sheds some light on how the development and establishment of microbiota take place on the human gut. For our next analysis we will apply appropriate statistics for twin studies and check specific species and their associations with host genetics.

Funding: CNPq, Fapesp

,,,,,,,,,,,,,,

Analysis of potential disease-causing variants in a patient with intellectual disability via whole-exome sequencing

Patricia de Cássia Ruy, Isabela Ichihara Barros, Reginaldo Cruz Alves Rosa, Jessica Rodrigues Plaça, Amanda Cristina Corveloni, Cibele Cardoso, Aline Fernanda de Souza, Carlos Alberto Oliveira de Biagi Junior, Ádamo Davi Diógenes Siena, Kamila Peronni Zueli, Maria Florencia Tellechea, Simone da Costa e Silva Carvalho, Greice Andreotti de Molfetta, João Pina, Wilson Araújo da Silva Jr

Faculty of Animal Science and Food Engineering of USP, USP, Brazil

Abstract

To realize the intellectual disability of a patient with none specific syndrome identified, the whole-exome sequencing (WES) technology was applied to search for potential disease-causing variants. Previous studies in our laboratory found a long noncoding RNA deleted in this patient and in vitro neural differentiation showed that patient cells are delayed in the differentiation process when compared to control cells. Besides the intellectual disability, this patient presents motor incoordination, flat feet, brachydactyly, obesity, macrocephaly, hypogenitalism and small mouth and teeth. To better understand this patient phenotype the WES of the patient and his family (father, mother and grandmother) were performed. DNA sequencing (Illumina TrueSeq Rapid Exome) of four blood samples was performed using an Illumina NextSeq 500. The quality of generated reads was obtained with FastQC software and the quality trimming was made by TrimGalore, considering a Phred Score of 30. The alignment of reads with the human reference genome (hg19) used BWA (Burrows-Wheeler Aligner) and the processing of alignment files was made by Picard tools. The variants (germinative mutations) were identified using Strelka. The possible pathogenetic significance of identified variants was assessed using Exomiser program with HPOs (Human Phenotype Ontologies) related to intellectual disability, autistic behavior, etc. This program integrates CADD (Combined Annotation Dependent Depletion), PolyPhen (Polymorphism Phenotyping), SIFT (Sorting Intolerant From Tolerant) and Mutation Taster to assist the variant impact. A filter of variants with maximum allele frequency of 2% was applied. An average of 122 million reads was generated per sample. The mean coverage was between 43x and 57x after removing duplicates. The analysis of the patient exome identified a total of 113, 200 variants. The top ten mutated genes are: CHR_START-U2 (440), Y_RNA (423), PRIM2 (274), CHR_START-AL592188.3 (256), bP-21201H5.1-IGHV1OR21-1 (252), RP11-96F8.1-KSR1P1 (233), snoU13 (213), BAGE2 (164), BX088702.2-CHR_END (147), CHR_START-Y_RNA (134). After annotation, filtering and prioritizing likely causative variants with Exomiser, the top 5 genes highlighted are: ATRX (2 missense X-recessive and 1 missense X-dominant variants), PTEN (1 splicing autosomal dominant and 1 splicing and 1 missense autosomal recessive variants), CHD7 (1 missense autosomal dominant and 2 missense autosomal recessive variants), AKT1 (1 missense autosomal dominant and 1 UTR5 autosomal recessive variants) and HDAC8 (1 missense variant X-dominant). A multi-sample Exomiser analysis will be made with the family samples to analyze the possible heredity of the identified variants.

Funding: CAPES, FAPESP, FAEPA, CNPq, FUDHERP

,,,

IDENTIFICATION OF FUNGI IN A BRASILIAN PAINT OF THE 20th CENTURY

Valquíria de Oliveira Silva, Paula Luize Camargos Fonseca, Maria Aparecida de Resende Stoianoff, Aristóteles Góes Neto

Universidade Federal de Minas Gerais

Abstract

The cultural heritage is our present time baggage and the heritage left to future generations. These cultural objects such paintings, sculptures, scrolls, archeologica sites are subject to several mechanismis of deterioration. These include microbial deterioration caused by fungi and bacteria that causes irreparable damage to cultural heritage. This work consisted in the isolation and identification of the fungi responsible for the deterioration in a twentieth-century Brazilian easel painting by italian-brazilian artist Lorenzato. The material for the isolation of the fungi was collected from the original work by means of sterile swabs in various regions of the pictorial surface between areas with fungal colonization and areas without apparent colonization. The samples were diluted in saline solution (0.85%). And by the serial dilution method, 100μL aliquots were obtained from the 10-1, 10-2, and 10-3 dilutions that were seeded on Potato Dextrose Agar (BDA) by the spread plate method. The isolated fungi were purified and observed at macroscopic and microscopic level for identification at genera level. We obtained 9 colonies of morphologically distinct fungi belonging to the following genera 2 Aspergillus, 4 Penicillium, 2 Hypocrea and 1 Nigrospora. Molecular characterization of the isolates was performed by extraction of fungal DNAs, PCR, and sequencing of the ITS4 and ITS5 regions of each sample. The results were processed in the Geneious software. They were analyzed on the Blast website for comparison and identification by similarity analysis with the NCBI Genbank database (nr) (the sequences deposited with the sequences obtained in this work were compared). The deteriorogenic species Aspergillus sydowii, Penicillium crysogenum, Hypocrea lixii and Nigrospora sphaerica were identified. The obtained results contribute for the understanding of the biodeteriogenic agents and also for the analysis of the conservation state of the studied object, allowing the adoption of mitigating measures in the scope of the preventive and curative conservation collaborating for the preservation of Lorenzato painting.

Funding:

,,,

THE PAINTING I LIVE: FUNGI IDENTIFIED ON PICTORIAL SURFACE

Valquíria de Oliveira Silva, Paula Luize Camargos Fonseca, Luiz Marcelo Ribeiro Tomé, Aristóteles Góes Neto

Universidade Federal de Minas Gerais

Abstract

The cultural heritage in their miscellaneous forms of artistic manifestation are subject to several mechanisms of chemical, physical, and biological deterioration. These biological mechanisms include microbial deterioration caused by fungi and bacteria. These cultural heritage refer to memory, cultural identity of a society, and represent the legacy left to future generations. The present study consisted in the identification of deteriogenic fungi present in the surface of the painting / panel of PORTINARI entitled "FREVO". This painting is very important because it was the last work of the italian-brazilian prepared by the artist before his death, and it is part of the Pampulha landscape Complex in BH / MG / Brazil, consisting of a set of assets listed by UNESCO as a cultural heritage of humanity. Samples were collected on the pictorial surface of the original work by means of sterile swabs in various regions between areas that had fungal colonization, and areas that did not have colonization. The samples were diluted in saline solution (0.85%), and by the serial dilution method, 100μL aliquots were obtained from the 10-1, 10-2, and 10-3 dilutions that were seeded on Potato Dextrose Agar (BDA) by the spread plate method. The isolated fungi were purified and observed at macroscopic and microscopic level for identification at genus level. Molecular characterization of the isolates was performed by extraction of fungal DNAs, PCR, and sequencing of the nrDNA ITS region of each sample. The results were processed in the Geneious software. They were analyzed on the Blast site for comparison and identification by similarity analysis with the NCBI Genbank database (nr). We found 14 fungal species from 8 different families: 6 Aspergillaceae from the following genera (4 Aspergillus and 2 Penicillium), 2 Didymellaceae whose genera are (Epicoccum and Didymella), 1 Apiosporaceae (Arthinium), 1 Hypocreaceae (Trichoderma), 1 Calosphaeriaceae (Pleurostoma), 1 Chaetomiaceae (Chaetomium), 1 Mytilinidiacea (Mytilinidion), and 1 Sporidiobolaceae (Rhodotorula). The obtained results contribute for the knowledge of the biodeteriogenic agents in the case studied the filamentous fungal and yeasts, and also for the analysis of the conservation state of the fine work, thus, allowing the adoption of mitigating measures in the scope of the preventive and curative conservation, collaborating for the preservation of the PORTINARI fine work.

Funding:

,,,,,

Comparative mitogenomics of Sugiyamaella species, yeasts of biotechnological importance

Paula Silva Matos, Heron Hilário, Rennan Garcias Moreira, Carlos Augusto Rosa, Thiago Mafra Batista, Glória Regina Franco

Universidade Federal do Sul da Bahia

Abstract

Microorganisms are widely used in the industry to produce several compounds. In the bioethanol production the most commonly used yeast is Saccharomyces cerevisiae, which converts sugars, mainly from sugarcane, into ethanol. However, this yeast is unable to degrade some plant polymers such as lignin to fermentable sugars, which prevents full use of the raw material. Some yeasts, like Sugiyamaella xylanicola, are able to metabolize these polymers, being promising candidates for second-generation ethanol production. These capabilities are conferred by enzymes encoded in the nuclear genomes of these microorganisms. However, phylogenetical information about fermenting yeasts can be obtained by sequencing their mitochondrial genomes, since they are small, ease to assemble and contain informative gene sequences. In this study we sequenced, assembled and annotated the mitogenome of the yeast S. xylanicola UFMG-CM-Y1884T, collected from decaying woods in the Caraça Mountains of Minas Gerais. The complete mitogenome of S. xylanicola was characterized and compared with previously published S. lignohabitans CBS 10342 and S. cerevisiae mitogenomes. The Sugiyamaella mitogenomes are 28 and 48 Kb long, and shorter than the 85 Kb S. cerevisae mitogenome. A total of 26 and 46 mitochondrial tRNAs were annotated for S. xylanicola and S. lignohabitans, respectively, which consist of more than the minimal set of 24 tRNAs required for translation, reported for S. cerevisiae. The Sugiyamaella mitochondrial genomes encode a complete respiration system, including NADH dehydrogenase complex I. In S. cerevisiae and several other yeasts, complex I is missing, which leads to a reduction of the respiratory energy yield. Further, we intend to investigate other aspects of mitochondrial genomes in order to discover additional possible Sugiyamaella advantages over the more commonly used yeasts for ethanol production.

Funding: Pronex/FAPEMIG

,,,,,,,

THE ROLE OF TUMOR HLA IN NON-MUSCLE INVASIVE BLADDER CANCER RESPONSE TO BCG IMMUNOTHERAPY

Ramon Torreglosa do Carmo, Giulia Wada Friguglietti, Diogo Bastos, Vitor Rezende da Costa Aguiar, Fabiana Bettoni, Diogo Meyer, Anamaria A. Camargo, Cibele Masotti

Abstract

The choice treatment for non-muscle invasive bladder cancer (NMIBC) is the complete transurethral resection of the tumor but followed by adjuvant immunotherapy with intravesical BCG (attenuated Bacillus Calmette-Guérin) instillations in high-risk cases of recurrence or progression. Immunotherapy significantly decreases the risk of disease recurrence because it eliminates tumor cells by stimulating the patient's immune response, capable of stimulating lymphocytes that supposedly recognize tumor antigens or the bacillus internalized by the tumor cells. Many patients do not respond to BCG treatment: 30-40% relapse and 10-25% develop muscle-invasive forms. There are no predictive biomarkers of BCG response implemented in clinical practice, and, in the context of immunotherapy with immune checkpoint inhibitors, there is a growing body of evidence pointing to predictive response factors related to the antigen presentation mechanism. HLA class-I molecules are directly involved in the presentation of neoantigens and, therefore, are also associated with tumor immune evasion mechanisms, including HLA somatic mutations, large deletions or transcriptional silencing. HLA polymorphism can also modify the response to immunotherapy, as it has been recently shown for melanoma and lung cancer: individuals with tumors bearing certain HLA-A and B supertypes or heterozygous for all HLA loci have lower risks of disease relapse and progression. In this work, we HLA-typed 35 primary tumors of NMIBC-BCG treated patients (17 responsive and 18 unresponsive to treatment) from exomes using Optitype. To estimate copy number variation (CNV) from exomes, we used CODEX. We investigated whether loss of diversity in HLA loci (homozygosity in at least HLA-A, B or C) or the presence B62 supertype was associated with survival rates. We did not observe a significant correlation between loss of diversity and relapse-free survival (RFS; Log-rank test p=0.52), neither progression-free survival (PFS; Log-rank test p=0.099), but individuals with tumors heterozygous for all HLA loci had longer RFS and PFS. We observed four tumors segregating the B62 supertype, but it was not significantly enriched in the unresponsive group (3 resistant: 1 sensitive, Fisher's exact test p=0.6041). We further evaluated if CNVs overlapping HLA-A, B and C could interfere in BCG response. We observed 1, 4, and 1 deletions and 1, 3, and 3 duplications overlapping HLA-A, B, and C coding regions, respectively, in 7 tumors. There was no correlation between response to treatment and CNVs overlapping HLA (Fisher's exact test p=1). Besides not finding a significant association with treatment response, which is attributed to lack of power for most of our analysis, it is of note that an unresponsive HLA-B homozygous tumor also had an overlapping deletion within the gene coding region, suggesting that the antigen presentation may be compromised in this case. This is the first evaluation of HLA variation in the context of NMIBC treatment, and we believed that an extended sample size may uncover the role of HLA in BCG-response.

Funding: Capes

””

Genomic and epidemiologycal analyses of Manhheimia haemolytica strains

Raquel Enma Hurtado Castillo, Janaína Canário Cerqueira, Rodrigo Profeta Silveira Santos, Marcus Vinicius Canário Viana, Vasco A de C Azevedo

UFMG

Abstract

Mannheimia haemolytica is a gram-negative bacterium, commensal and opportunistic pathogen, and the primary agent of respiratory infections on ruminants. This bacterium causes bovine respiratory disease, a disease that generates a great economic loss in the cattle industry. Some pathogenic strains are associated to a specific serotype and the presence of integrative conjugative elements (ICEs) containing multi-drug resistance genes. To characterize the genomic features, pathogenesis, and distribution of this specie, we performed a genomic and epidemiological analyses of 113 M. haemolytica strains from diverse hosts, mostly cattle, and clinical or non-clinical status. The serotypes were classified as serotype 1 (48.67%), 2 (31.85%), 6 (13.27%) and unknown (6.19%). According to Multilocus Sequence Typing (MLST), the strains were classified as ST1 (61.94%), ST2 (30.08%), ST7 (2.26%), ST47 (1.13%), ST3 (1.13%) and unknown (3.39%). The pangenome was estimated as open. Phylogenetic analysis using whole-genome single nucleotide polymorphisms (SNPs) showed that most strains with the same serotype were clustered together. Principal Component Analysis (PCA) based on the accessory genes segregated all identified serotypes. The serotypes 1 and 6 are discretely segregated, but all strains belong to sequence type ST1, while all serotype 2 strains belong to sequence type ST2. Strains with unknown serotype did not form clusters. The integrative conjugative elements ICEPmu1 and ICEMh1 were present only in strains from USA, in both clinical and non-clinical sample, but not in all strains. The results suggest that the in silico-identified serotypes could be discriminated by SNP. In addition, the serotype 2 could be differentiated from serotypes 1 and 6 by the accessory genome. The association of genomic features with clinical or non-clinical isolates, and host species could not be evaluated due to lack of available data. Integrative conjugative elements were reported only in USA isolates, but were not specific to clinical or non-clinical isolates. Virulence factor distribution across strains will be performed in further analyses. Our findings identified genomic features that could be associated to serotypes and geographical location that could help to develop strategies for surveillance, control and prevention on respiratory infections by M. haemolytica.

Funding:

,,,

Assessment of fecal microbiome differences in captive and non-captive howler monkeys: implications for conservation planning and management

Raquel Riyuzo de Almeida Franco, Gustavo Ribeiro Fernandes, João Carlos Setubal, Aline Maria da Silva

Universidade de Sao Paulo

Abstract

Howler monkeys (Alouatta spp) are endemic of South American tropical forests and are highly susceptible to the yellow fever virus, thus playing an important role as sentinels of outbreaks. Brazil is experiencing its worst yellow fever outbreak in decades, which is wiping out some wild howler monkey populations. One strategy to regrow endangered populations is the introduction in the wild of captive - bred animals. Due to the significance of the gut microbiome in animal health, understanding its composition and function may help conservation planning and management of these species. The present study aims to compare the gut microbiomes of captive and non - captive howler monkeys. Fecal samples from Sao Paulo (Brazil) Zoo Park captive animals and non - captive animals that live in the park's Atlantic rain forest patch were obtained in two different seasons in two consecutive years, followed by 16S amplicon and shotgun sequencing. Data analysis revealed differences in the microbial community structure, diversity, and function between the two populations, with non - captive individuals showing higher phylogenetic diversity indices and enrichment of specific metabolic functions. The microbiome of non - captive monkeys appears to be more susceptible to seasonal changes than the microbiomes of captive individuals, perhaps due to seasonal changes in food availability. In the microbiota of captive animals we did not identify members of the genus Faecalibacterium, which was identified in the non - captive samples, and which is an abundant genus in the healthy human gut microbiome. Several novel bacterial and viral genomes were recovered from the shotgun sequences.

Funding: Fapesp, Capes, Cnpq

,,,

A STRUCTURAL AND EVOLUTIVE APPROACH ON NUCLEOTIDE EXCISION REPAIR IN EUKARYOTES

Rayana dos Santos Feltrin, Ana Lúcia Anversa Segatto, Tiago Antonio de Souza, André Passaglia Schuch

Universidade Federal de Santa Maria

Abstract

Among several DNA repair mechanisms developed throughout evolution, the nucleotide excision repair (NER) is the most versatile repair pathway as it removes a wide range of structurally unrelated lesions that distort the double-helix. Given its importance in the maintenance of genome integrity, it appeared earlier in the evolution of species. However, few current studies involving NER have an evolutive approach. Moreover, the availability of a large amount of data on genome databases makes possible to retrieve sequences of NER components from many species. Therefore, we performed a homology assessment for ten main proteins from the NER pathway in 13 eukaryotic organisms (Mus musculus, Gallus gallus, Alligator mississippiensis, Xenopus laevis, Danio rerio, Caenorhabditis elegans, Drosophila melogaster, Arabidopsis thaliana, Trypanosoma cruzi, Giardia lamblia, Saccharomyces cerevisiae, and Schizosaccharomyces pombe). To do so, we used human reference sequences as queries to perform a tblastn search against each genome, aiming to obtain the nucleotide and respective protein sequences from each species. In order to get the best orthologue candidates, we applied four criteria as sequence filters: e-value, percent identity, protein domains in common, and protein sequence length. To identify the conserved protein domains, we submitted the sequences to Simple Modular Architecture Research Tool (SMART), considering Pfam database. The protein sequences from each NER component were aligned on MEGA 7 software by MUSCLE, and then the most conserved regions were obtained with Gblocks 0.91b. To set the best evolutionary models, the resulting alignments were evaluated on ProtTest 3.4.2. Thus, an estimate of the evolutive distance of the retrieved sequences in relation to the human ones was conducted on MEGA 7. We also run a phylogenetic reconstruction for each NER protein by using the Neighbor-Joining algorithm, considering the substitution model JTT+G, 1000 bootstrap replicates, and pairwise deletion with the same software. Additional analyses were carried out to assess the number of introns of the genes by using Gene Structure Display Server 2.0 (GSDS 2.0), as well as gene sizes, relying on information from GenBank. In general, our results show that most of NER components analyzed have homologs in the eukaryotic species, but some of them were not detected using our criteria. We also observed that although several variations are found in the gene structures, the protein structure has only small changes. Furthermore, this work suggests further research to better investigate the real biological efficiency of the NER pathway in some eukaryotic organisms.

Funding: CAPES and CNPq

,,,,,,,,,,,,,,,,,,,,

Whole-exome sequencing evaluation of BCG responsiveness in high-risk non-muscle invasive bladder cancer (NMIBC)

Diogo A Bastos, Romulo L Mattedi, Rodrigo Araujo Sequeira Barreiro, Filipe Ferreira dos Santos, Vanessa Candiotti Buzatto, Cibele Masotti, Jussara M Souza, Mariana Zuliani Theodoro de Lima, Giulia W. Frigugliett, Carlos Dzik, Denis L Jardim, Rafael Coelho, Leopoldo A Ribeiro Filho, Mauricio D Cordeiro, William C Nahas, Evandro S de Mello, Roger Chammas, Luiz Fernando Lima Reis, Fabiana Bettoni, Pedro Galante, Anamaria A. Camargo

Translacional Oncology Center, ICESP

Abstract

Non-Muscle Invasive Bladder Cancer (NMIBC) accounts for 70-80% of the cases of bladder cancer. The gold-standard therapy for NMIBC are a transurethral resection of the lesion and an intravesical injection of Bacillus Calmette-Guerin (BCG). Immunotherapy using BCG are associated with the reduction of tumor recurrence and progression, but only approximately 50% of the patients benefit from this therapy and approximately 20% of the BCG treated patients interrupt the therapy due to side effects. The underlying mechanisms associated with the response of BCG immunotherapy are not yet well understood and there is not an available biomarker for response. Here, we sequenced the whole exome (WXS) from tumor of 35 (17 responsive, BCG-R; and 18 unresponsive, BCG-UR) high-risk NMIBC patients from Instituto do Câncer do Estado de São Paulo (ICESP) and performed a variant calling pipeline in order to find genomic variables associated with patient outcome. For the variant calling, we used the GATK best practices for variants calling and thereafter we filtered out germline variants by the presence in populational variants database (ExAC and 1000 Genomes) and the recurrence in our cohort. Our results show differences of tumor mutational burden (TMB) between BCG-R and BCG-UR (p-value = 0.045), in which the low-TMB group showed a higher relapse-free survival than the high-TMB group (p-value = 0.0092). We also evaluated tumor heterogeneity by Mutant-Allele Tumor Heterogeneity (MATH) score, however no statistical significance was found between BCG-R and BCG-UR. In the end, we found an import result to non-muscle invasive bladder cancer patients: TMB as a potential predictive biomarker for BCG immunotherapy response.

Funding: CNPq

,,,,,,,,,

Comparative genomics analysis and classification of the Lactobacillus casei species

Rodrigo Bentes Kato, Diego Lucas Neres Rodrigues, Juan Luis Valdez Baez, Roselane Gonçalves dos Santos, Stephane Fraga de Oliveira Tosta, Alessandra Lima da Silva, anne cybelle pinto gomide, Alfonso Gala-Garcia, Francielly Rodrigues da Costa, Vasco A de C Azevedo

UFMG

Abstract

Lactobacillus taxon has been widely studied for its probiotic characteristics, including immunomodulatory and metabolic properties. The World Health Organization (WHO) points out that the genus and species classification of any probiotic is crucial before its use, genomic studies of lactic bacteria are of fundamental importance in the legal, biological and safety aspects. Therefore, gene blocks related to pathogenicity mechanisms should be studied to ensure the use of bacteria in the standards provided by WHO, although the literature indicates that some are symbiosis mechanisms. The aim of the present study was to compare the genome sequence of six strains of Lactobacillus casei and to show possible divergence points capable of segregating them from the others in the genus. The six complete L. casei genomes were obtained from the platform (NCBI) and standardized for annotation using the PROKKA software. For the comparison was developed a pipeline based on the Average Nucleotide Identity (ANI), performed on the JSpeciesWS platform, and on the construction of a phylogenomic tree of individuals of the genus using the PATRIC platform. Our analysis of genomic plasticity was performed using the visualization tools GIPSy and BRIG. We obtained images showing significant divergent points between the genomes applied to the study, as well as a low nucleotide identity. Islands of pathogenicity totally or partially shared among all genomes were detected. This last result suggests that the use of this bacterium is not safe to use as a probiotic. There is still great difficulty in classifying L. casei at the species level. As phylogenetically, L. casei is still very confused within the genus as other family members, which negatively influences its taxonomic identification which is crucial as a safety measure for its use in food.

Funding: CNPq, CAPES, FAPEMIG

,,,,,,,

Genomic characterization of Lactobacillus delbrueckii CIDCA 133: a potential probiotic strain

Rodrigo Profeta Silveira Santos, Luís Cláudio Lima de Jesus, Marcus Vinicius Canário Viana, Janaína Canário Cerqueira, Mariana Martins Drumond, Pamela Mancha-Agresti, Bertram Brenig, Vasco A de C Azevedo

University Göttingen

Abstract

The probiotic potential of Lactobacillus delbrueckii CIDCA 133, isolated from raw cow milk, have been validated by in vitro studies that showed its resistance to high acid and bile concentrations, the ability to inhibit the growth of pathogenic microorganisms such as enterohemorrhagic Escherichia coli, resistance to antimicrobial peptides of erythrocyte cells, and anti-inflammatory and immunomodulatory properties. However, despite of these important traits, little is known about the molecular mechanisms involved in these processes. In this study, we performed a comprehensive analysis of its genome to better understand the molecular basis of its in-vitro-tested probiotic characteristics. The genome of L. delbrueckii CIDCA 133, herein named CIDCA 133, was sequenced using Illumina HiSeq platform and assembled using Spades. The assembly resulted in 70 contigs and a 6223 bp sized cryptic plasmid. The characterization of its genome confirmed that CIDCA 133 belongs to the subspecies lactis, with 98.21% of ANI compared to the type strain L. delbrueckii subsp. lactis DSM 20072. Comparative genomics analyses were conducted including CIDCA 133 and 64 publicly available L. delbrueckii genomes. Despite the expected clustering of the subspecies in the phylogenomic tree, generated by the software PEPR, divergent identifications were found for some L. delbrueckii subsp. bulgaricus strains, which grouped in the same clade with another L. delbrueckii subsp. lactis. Given the similarity between the subspecies, our results suggest that some genomes have been misclassified in the GenBank. Hence, greater care should be taken with their characterization, using genomic analysis as part of this process. Preliminary functional analysis shows that CIDCA 133 has 18 exclusive genes, one of these (nagB) required for the Amino sugar and nucleotide sugar metabolism pathway. These genes might be related to the metabolic pathway associated with the singular probiotic characteristics described for this strain, and thus, more analyses will be performed in order to characterize them.

Funding:

,,,

The rare lncRNA GOLLD is widespread and structurally conserved among Mycobacterium tRNA arrays

Sergio Mascarenhas Morgado, Deborah Antunes, Ernesto Raul Caffarena, Ana Carolina Paulo Vicente

Oswaldo Cruz Institute

Abstract

Noncoding RNA (ncRNA) genes produce transcripts involved in catalytic or regulatory functions, some of them presenting highly complex structures. GOLLD RNA is the third-largest bacterial ncRNAs known (800 bp); however, its function is still unknown. The GOLLD RNA gene is generally found associated with tRNA genes and supposed to be chromosome- and phage-encoded in bacteria from Lactobacillales and Actinomycetales orders. Besides, the only inferred GOLLD RNA structure was mainly based on metagenomic sequences. To explore GOLLD in bacterial genomes, we mined GOLLD gene in thousands of Mycobacterium and virus genomes using Infernal software, identifying it in 350 mycobacteria (including two in megaplasmids) and 39 virus genomes, mainly associated with tRNA arrays. Mycobacterium GOLLD genes are highly diverse and distributed in three clades: Mycobacterium exclusive; Mycobacterium and mycobacteriophages; and mycobacteriophage exclusive. We also determined the secondary structure of each clade using R2R software based on GOLLD alignments generated by Infernal software. All clades displayed a 3′ half conserved structure including utter E-loops pseudoknots substructures, also shared by non-Mycobacterium GOLLD while the 5′ half motif was different among the clades. In some cases, an ORF, coding a tRNA or transposase gene, was predicted in the 5′ half motif. Moreover, in vitro assays determined the expression of GOLLD RNA gene present in a plasmid harbored by a Mycobacterium isolate from Atlantic Forest soil. Our study showed that the long ncRNA GOLLD is widespread within Mycobacterium in association with tRNA arrays, besides strengthening its structure, previously predicted based in metagenomic sequences.

Funding: Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq); Coordenação de Aperfeiçoamento de Pessoal de Nível Superior -Brasil (CAPES); Oswaldo Cruz Institute

””

Beginning and end of composting as viewed through metagenome-assembled genomes

Suzana Eiko Sato Guima, Roberta Verciano Pereira, Layla Martins, Aline Maria da Silva, João Carlos Setubal

Programa Interunidades de Pós Graduação em Bioinformática, Universidade de São Paulo (USP)

Abstract

Composting is a process carried out by many different microorganisms capable of biomass degradation. In this work our goal is to analyze this microbial diversity focusing on metagenome-assembled genomes (MAGs) recovered from the beginning and the end of the composting process. We collected five samples from inoculum (or decompost) and two from mature compost in the composting facility at the São Paulo Zoo Park. Decompost consists in a compost pile in later phase of the composting process that is sampled and added with alternate layers of plant and animal residues to start a new composting pile. Mature compost is the product of the composting process ready to be used as fertilizer. Sequencing of total DNA was done with Illumina technology. Decompost reads were co-assembled using metaspades. We used metawrap pipeline for MAG recovery with Metabat2, MaxBin2, and Concoct as chosen binners. The same method was used to recover MAGs from mature compost. MAGs were classified using GTDB-tk. Quant_bins from metawrap was used to quantify the abundance of MAGs against time-series composting data (ZC4 compost pile) from a previous study. 12 out of 17 MAGs recovered from decompost were classified as Firmicutes. For mature compost, 8 out of 16 MAGs were classified as Actinobacteria. Abundance over time of MAGs mapped against reads from time-series samples exhibited a few trends: 1) some Firmicutes recovered from decompost in this study tends to present a high abundance in the beginning of the composting process and a slight increase after turning procedure; 2) some Actinobacteria recovered from mature compost tend to increase in abundance right before the turning procedure and during the final phase of the composting process. This variation suggests microorganisms presenting the trend 1 are favored by easily degradable organic nutrients and oxygen available in the beginning of the composting and after the turning procedure. As oxygen and easily degradable nutrients become scarcer and difficult to access, microorganisms able to degrade the remaining lignocellulosic materials are selected positively, showing the trend 2. Other analysis is ongoing for further understanding of the composting microbiology.

Funding: CNPq, FAPESP, CAPES

,,,,,,,

Comparative genomic analysis of Corynebacterium pseudotuberculosis: A quest for biofilm biosynthesis genes.

Thiago de Jesus Sousa, anne cybelle pinto gomide, Letícia de Castro Oliveira, Nubia Seyffert, Bertram Brenig, Mateus Matiuzzi Costa, Siomar de Castro Soares, Vasco A de C Azevedo

University Göttingen

Abstract

Corynebacterium pseudotuberculosis is the etiological agent of Caseous Lymphadenitis (CLA) in sheep and goats (C. pseudotuberculosis biovar ovis), and present some cases of infections in large ruminants as well. Our laboratory studies C. pseudotuberculosis as a model organism and, in 2013, Soares et al. performed a pan-genomic study of C. pseudo-tuberculosis using fifteen strains. After that, a hundred and eight strains were sequenced and deposited in the National Center for Biotechnology Information (NCBI), and several studies such as Exoproteome, Transcriptome, and Proteome were done. Now, we want to associate these studies and explore other areas of pan-omics analyses. In this work, we propose to research evidence of genes belonging to the biosynthesis of biofilm in C. pseudotuberculosis genomes. For this, thirty C. pseudotuberculosis samples from lymph nodes of abscesses of goats and sheep from Pernambuco-Brazil were isolated. All strains belonged to the biovar ovis group, proven by the nitrate reductase biochemistry test and multiplex PCR. These strains were sequenced by the Illumina Hiseq paired-end platform with 450bp insert, and 151bp per reads length. All strains have an average genome size of 2, 34Mb, the genomes were de novo assembled using SPAdes v. 3.9.1 software with 121 k-mers, and we got five or six contigs per genome with coverage of 971.40 fold. After, all the thirty genomes were deposited in the GenBank database in NCBI. For comparative genomic analyses, we used the programs Gegenees v. 3.1, BRIG v. 0.95, MAUVE using progressiveMauve and ACT: the Artemis Comparison Tool. Previous results described that among the thirty C. pseudotuberculosis samples, CpCap3W and CpOVI03 strains showed characteristics of biofilm formation, different from CpOvi2C and CpCAPJ4 strains. From the results of comparative genomics, the strains are 99.5% similar and very clonal. We could not find a concise difference about the presence or absence of genes be-longing to biofilm biosynthesis among the studied strains. However, in the Cp38MAT strain, a unique region with a cluster of eight genes was found. Among these genes, we identified the dps gene (DNA starvation/stationary phase protection protein) involved in iron homeostasis, in addition to 5 proteins involved in the process of transport and plas-matic membrane structure, and VanZ gene (Glycopeptide antibiotics resistance protein). Probably, this metabolic process of biofilm biosynthesis may be better clarified through specific transcriptome and proteome studies.

Funding: Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq)

,,,,,,,,,

Studies of LDL receptor activity in patients with familial hypercholesterolemia

Thais Kristini Almendros Afonso, Victor Fernandes de Oliveira, Glaucio Monteiro Ferreira, Jéssica Bassani Borges, Gisele Medeiros Bastos, Profa. Dra. Tania Cristina Pithon-Curi, Renata Gorjão, Rui Curi, Rosário Dominguez Crespo Hirata, Mario Hiroyuki Hirata

Laboratory of Molecular Research in Cardiology (LIMC), Dante Pazzanese Institute of Cardiology, São Paulo, Brazil

Abstract

The primary and more frequent alteration of the Familial Hypercholesterolemia (FH) affects the LDL receptor gene (LDLR), with more than 1600 mutations described. These alterations result in the compromise of decreased LDL removal and consequently their accumulation in plasma. Nowadays, genomic ultrasequencing technology generated high throughput results, thus, it is necessary to evaluate in vitro functional activity to make an association with disease. The aim of the present study was to analyze LDL receptor functional activity, in mononuclear cell collected from patients that were identified new variants in LDLR gene that showed correlation with clinical manifestation of hypercholesterolemia after in silico analysis of protein structure. To meet this goal, the T lymphocytes from the thirty FH carriers were isolated from the peripheral blood, cultured and challenged for the expression of LDL receptors, incubated with labeled LDL for binding assessment and internalization by the cells of each patient. The LDLR variants were selected after exome sequencing of a panel of 61 genes related to cholesterol homeostasis using the MiSeq platform System (Illumina). The in silico analyses were performed through a specific pipeline for variant calling and annotation. Initially, these sequences were aligned to the reference human genome (GRCh38) using BWA. Afterwards, GATK 4.0 was used for base quality score recalibration and variant calling to identify high-quality SNVs. Finally, annotations and functional effect predictions for the SNVs were performed by PolyPhen-2, SIFT, PROVEAN and ANNOVAR. In order to elucidate the internalization of LDL particles, in silico studies were performed to evaluate the functional impact of seven LDL receptor variants with APOB. For this, 3D models were built by homology modeling to obtain a predictive structure. Thus, Protein-protein docking was performed on the server ClusPro 2.0, using a standard set up, between LDLr (PDB: 1n7d) and APOB (PDB: 1lsh). In this context, in vitro study it was possible to note an increase in both the mean fluorescence of binding and binding and internalization in relation to the amount of LDLr on the cell surface. Corroborating to in vitro studies and molecular docking, variant p.Gly592Glu (rs137929307) suggest an increase complex formation LDLr-APOB (LDLr: 31.5% and LDL endocytosed: 50.7%) while p.Asp601His (rs753707206) suggest less affinity between LDLr and APOB (LDLr: 98% and LDL endocytized: 37.4%). Therefore, understand related mechanisms can measure the functional impact correlated with the pathogenicity of LDLr variants in HF.

Funding: FAPESP - Project: 2018/11917-6

”

Resistome analysis of bacterial genomes from bloodstream infection reveals antibiotic efflux as the main resistance mechanism in blood isolates.

Willian Klassen de Oliveira, Luis Gustavo Morello, Helisson Faoro

Instituto Carlos Chagas - FIOCRUZ PR

Abstract

Nowadays, bacterial resistance to antibiotics is a global public health problem and studies point to infections caused by resistant bacteria becoming the leading cause of death by 2050. The phenomenon of resistance is even more severe when associated with blood infection that may evolve and become sepsis. A deeper understanding of what are the main pathogens and the resistance they have is essential to overcome this threat. For this purpose, in this work we conducted a large meta-analysis study with 3, 872 genomes of bacteria isolated from bloodstream infections. Through search on HMM models of proteins proved to be involved in bacterial resistance to antibiotics, we identified 71, 675 resistance proteins, which were classified according to their mechanism of action, proteins types and class of antibiotic to which it gives resistance. We also analyzed the diversity of bacterial taxons in these samples. We found a prevalence of Proteobacteria and Firmicutes phyla representing respectively 56, 2% and 41, 2% of the total organisms. In general, the main genera present in this study were Staphylococcus and Klebsiella. These genera also have high numbers of resistance genes, with an average of 26 and 30 resistance genes per genome, respectively. Another genus that stand out in the study is Elizabethkingia, appearing as a potential emerging pathogen. Despite having few occurrences in the sample, it has 17 resistance genes per genome on average, a high number when compared to other genus. Through our analysis the main mechanism of antibiotic resistance in bloodstream infection is that related to antibiotic efflux with 72.7% of the classified proteins in this category. At taxonomic levels, we found differences between the resistance mechanisms in gram-positive and gram-negative bacteria. The RND type transporters are more present in Proteobacteria, while in Firmicutes the ABC transporter is the most used to export antibiotics. The classes of antibiotic to which bacteria presents the most resistance genes in all phyla is beta lactams and aminoglycosides. The identification of antibiotic resistance mechanisms in bacteria from bloodstream infection is fundamental as it may contribute to the development of treatments and strategies to combat these bacteria.

Funding: CAPES

,,,,,

An In Silico approach for the identification of vaccine and drug targets against Mycoplasma genitalium, causative agent of sexually transmitted pelvic inflammatory disease (PID)

Arun Kumar Jaiswal, Wylerson Nogueira, sandeep tiwari, Rommel Thiago Jucá Ramos, Vasco A de C Azevedo, Siomar de Castro Soares

Universidade Federal do Triângulo Mineiro

Abstract

Mycoplasma genitalium is a sexually transmitted pathogen characterized as a pleiomorphic, flask shaped, slow growing and obligate intracellular bacterium. It is one of the STI (sexually transmitted infections) pathogens associated with non-gonococcal urethritis in men and several inflammatory reproductive tract syndromes in women such as cervicitis, pelvic inflammatory disease (PID) and infertility. Some studies have reported the infection of M. genitalium as a cause for infertility and adverse pregnancy outcomes such as preterm labor. Currently, treatment of most M. genitalium infections occurs mainly in the context of syndromic management for urethritis, cervicitis, and PID, owing to the lack of diagnostic test availability. With the advent of new high-throughput sequencing technologies and the rise of genomic data, scientists are able to use computational methods to identify new targets, which are time and cost effective in compare to classical approaches. Reverse vaccinology (RV) and subtractive genomics are conventional and popular approach in the post-genomic era for the prompt identification of novel vaccine and drug targets. In this study, the prediction of putative vaccine and drug targets against Mycoplasma genitalium, using reverse vaccinology and subtractive genomics is carried out. We used 10 strains of Mycoplasma genitalium for comparison. Briefly, we used a combined reverse vaccinology and subtractive genomics approach and identified 12 putative antigenic proteins as vaccine targets and 7 drug targets. Furthermore, the molecular docking analysis was performed with 5000 antimicrobial natural compounds downloaded from ZInc database. The drug-like natural compounds showed the most favored binding affinity against predicted drug targets, which can be a candidate therapeutic target in the future against M. genitalium. We hypothesize that these identified therapeutic targets and antimicrobial drugs could be considered for prophylaxis of M. genitalium and hence should be subjected to further experimental validations.

Funding: CAPES, FAPEMIG, CNPq

”

Possible bias in predicting essential genes

Zandora Celeste Hastenreiter Ferreira Nunes, Francisco Pereira Lobo, Giovanni Marques de Castro

Universidade Federal de Minas Gerais

Abstract

The ever-growing amount of complete genomes calls for the creation of computational strategies in order to extract biologically meaningful information from this data. The prediction of essential genes is one of the areas where there is a growing effort to create such strategies, since they are attractive targets for intervention in parasites, vectors and pests. The use of machine learning can help detecting such genes and reduce the burden for experimental validation, aiding the choosing of interesting candidates. For this work, we used a Random Forest algorithm to predict essentiality in insect genes, as this information would potentially help in the development of specific, low impact bioinsecticides. Most of the currently available software for this task relies on information based on expensive experimental data, such as gene annotation, gene expression profiles and interaction networks. Our approach is homology-independent in the sense that the features are calculated from its nucleotide and peptide sequence data alone, such as amino acid/nucleotide frequencies and sequence entropy. Initially, we recovered alleles of Drosophila melanogaster from FlyBase classified as loss of function, amorphic and hypomorphic. For the essential genes, we recovered the alleles annotated as having a lethal phenotype and, for the nonessential genes, we recovered those that are not annotated as lethal. We converted the allele ids to gene ids and retrieved the longest isoform nucleotide sequence for that gene. The training dataset consisted of 1256 essential genes and 636 nonessential genes, and the test set consisted of 47 essential and 88 nonessential genes. The genes present in the test dataset were curated from published works. The model consisted of 1000 trees and 10 fold cross-validation, repeated 5 times. We obtained an AUC of 0.7238. We found this result to be caused by a bias introduced by a small group of evolutionary old essential genes and young nonessential genes; when we withdraw these groups from the test set, the AUC dropped to 0.578. Surprisingly, when we use only this biased group as the test set, the AUC obtained was 0.995, suggesting our classifier learned to discriminate between old and young genes. As future work, in order to further evaluate our model, we will retrieve orthologs present only in D. melanogaster and D. simulans and consider these as young genes; orthologs also present in D. mojavensis and D. virilis will considered old genes. The new test set will then comprise essential young genes and old nonessential genes.

Funding: CAPES, CNPq, PPG Genética - UFMG

5 — Phylogeny and Evolution

,,,,,,

Positive selection evidences on Moniliophthora PR-1 genes suggest evolution towards pathogenicity role

Adrielle Ayumi de Vasconcelos, Renata Baroni, Paulo M. Tokimatu Filho, Paulo J. P. L. Teixeira, Marcelo Falsarella Carazzolle, Gonçalo Amarante Guimarães Pereira, Juliana José

Institute of Biology; UNICAMP; Brazil

Abstract

Moniliophthora perniciosa is a basidiomycete fungus with three known biotypes, being the C-biotype the one that infects Theobroma cacao, causing witches broom disease (WBD).The arrival of WBD in cocoa plantations in Bahia led to enormous economic and social damage. In order to aid the development of forms of pathogen control, the understanding of the plant defense mechanisms and virulence factors of M. perniciosa is necessary. Pathogenesis-related 1 (PR-1) proteins, which belong to the SCP/TAPS or CAP superfamily, are widespread markers of the induced defense response in plants against pathogens. Interestingly, M. perniciosa's genome contains 11 PR-1-like genes (named MpPR-1a to k), many of them being highly up-regulated genes during WBD. In this study, we carried out the evolutionary analysis of M. perniciosa's PR-1 genes, also searching for evidence of positive selection shaping those proteins. We used putative PR-1 gene families identified across the genomes of 22 Moniliophthora isolates (18 M. perniciosa and 4 M. roreri) for inference of the gene phylogenetic history within Moniliophthora and also for phylogenetic inference among orthologous PR-1 identified from other 17 species from Agaricales order. These analysis revealed that PR-1c, a highly expressed gene during WBD, is possibly exclusive to C-biotype isolates and is a recent paralog of PR-1j. Besides, the five most recent PR-1 genes in the gene phylogenetic tree are the ones expressed during infection (f, g, i, k, h). The phylogenetic inference with Agaricales PR-1 genes revealed that PR-1a, b, d, and j, were the gene families with most orthologous from the various species, while PR-1i, g and k are only represented in Moniliophthora isolates. PR-1a, b and d are ubiquitously expressed during M. perniciosa mycelium stages and PR-1j is expressed in basidiomata, suggesting that these proteins might have a role for fungi basal metabolism. Evolutionary models of positive selection were tested in all Moniliophthora PR-1 gene families using the dN/dS ratio with codeml package of PAML4. While branch-sites model using C-biotype branches as foreground did not detect evidence of positive selection, testing for sites model detected signals of sites under positive selection in four PR-1 families (PR-1f, g, h, i), which are strongly expressed during WBD. These proteins might have evolved from non-pathogenic PR-1 proteins and became advantageous for the pathogen's success in host infection during Moniliophthora recent evolution, revealing important proteins and codon-targets for the pathogenicity of M. perniciosa in cocoa.

Funding: Supported by FAPESP (2017/13015-7)

,,,

An integrative approach to understand species delimitation in Petunia

Ana Lúcia Anversa Segatto, Maikel Reck-Kortmann, Caroline Turchetto, Loreta Brandão de Freitas

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

Abstract

Plants are considered to be plastic organisms, not only due to the phenotypic plasticity they usually present but also because they can hybridize frequently. Thereby, it is difficult to determine if the morphological variation seen, in what is considered one plant species, is morphological plasticity, hybrid organisms, or local adaptation. Taking this into account, simulated datasets have shown that the most used methods of species delimitation are negatively influenced by gene flow and incomplete lineage sorting. The long corolla tube clade of Petunia, formed by Petunia exserta, Petunia secreta and three subspecies of Petunia axillaris, presents taxa with morphological variation, genetic sharing, and disjoint distribution that make it a good system to study the effects of population phenomena in species delimitation. In this work, our objective was to clarify the evolutionary relationships among long corolla tube taxa, including canonical, non-canonical and geographic disjointed individuals within each taxon. To do this, we sequenced eight nuclear regions, five plastid DNA markers, and genotyped seven microsatellite loci. The phylogenetic relationships among the lineages were estimated using the Bayesian Inference as implemented in BEAST 1.7, and species delimitation was conducted using the Program BPP 3.4. The software Structure 2.3 was used to perform clustering analysis based on microsatellite data. The supergene tree showed better resolution than the species tree. Petunia species previously described as belonging to the long tube clade formed a monophyletic group in the supergene tree and were distributed in two main subgroups. Multiple runs of BPP with different MCMC parameters and guide trees gave similar results; the posterior probabilities values were higher considering eight species. In contrast, microsatellite markers analyses indicated the occurrence of two genetic components. The long corolla tube clade of Petunia encompasses markedly different taxa, regarding morphology and life habits. The multigene and multimarker approaches used here to disentangle the evolutionary relations among these taxa confirmed their genetic identity; however, it did not agree with the current taxonomic classification. The observed scenario possibly involves a complex interaction of different environmental, phenotypic, and genetic phenomena, which is similar to what is proposed to a great number of species. Understanding the relationship between genetic diversity and other variability sources is extremely important to the preservation of evolutionary lineages, mainly in front of the global environmental changes, and can contribute to understanding how the taxa identity is maintained. Incongruence between different data sources may be the key to understand the evolutionary history and new Bioinformatics approaches are necessary to deal with that.

Funding: CNPq, CAPES, PPGBM-UFRGS.

,

Retina development pathway construction and evolutionary analyses through text-mining and orthologue clustering tools.

Arthur Pereira da Fonseca, José Miguel Ortega

Universidade Federal de Minas Gerais

Abstract

The mammalian retina is composed of six major types of neurons and one type of glia cell. During retinogenesis, all types of cells are specified from a multipotent progenitor pool. Although several transcriptional factors have been documented to play a role in the differentiation, we lack a study describing the development's full pathway. Therefore, understanding how these factors interact is essential to develop and improve medical treatments such as tissue and cells transplants. For this work we used bioinformatics text-mining tools to reveal the interactions between the genes involved in retinogenesis. First was made a list of abstracts with a PubMed query search for "retina development". Then we used the Medline ranker tool, with six selected articles as a training set, to rank in this list the top thousands abstracts to use in PESCADOR. After this we used the platform PESCADOR to search and highlight the biointeractions described in those abstracts. The genes in the final pathway were also run in another program, TaxOnTree, a tool to create homologous clusters along with a taxonomic classification. The information about in which taxon it was first detected, allowed us to determine the Last Common Ancestral for each gene. With these tools we found 69 genes implicated in the retina's development regulatory pathway. And the cluster analyses revealed that the great majority have their origin in Euteleostomi, the clade that groups fishes and men. Only one gene (CRX) was found to have its origin within the mammalian class. It interacts with other transcription factors, such as NRL, from Tetrapoda, RORA, from Euteleostomi, and RAX, from Gnathostomata, to regulate the transcription of photoreceptor specific genes, including the Opsin family, from Euteleostomi. Making it a key element on the differentiation and viability of cone and rod photoreceptors. The use of text-mining tools allowed us to combine together several information for a better visualization of a biological process. Also, the homologous cluster analyses enabled a better understanding about the evolutionary process, and to extend this visualization for other organisms. In conclusion, with these tools we were able to construct and analyze the retinogenesis pathway and probable origin of their genes. Evolution of retina's development appears to occur with the appearance of fish and some key components of the human pathway are restricted do mammals.

Funding: CAPES Computational Biology Networks: Biologia Sistêmica do Câncer, BSC.

”

If menstruation is recent in evolution of great primates, how ancient are the genes that control its periods?

Andre Luiz Garcia de Oliveira, Arthur Pereira da Fonseca, José Miguel Ortega

Universidade Federal de Minas Gerais

Abstract

Orthologues grouping provides a better knowledge of the sequences producing a certain function, but allow us to verify the taxonomic distribution and, after that, infer the clade of origin of novel functions. With this in mind we set up to study the origin of genes implicated with menstruation. Menstruation is a rare event in mammals, but it is an essential phase for the human reproductive cycle of some great primates in the ancient world. In them, ovarian steroid hormones regulate endometrial function and human menstruation. In humans, after ovulation, the corpus luteum secretes high levels of progesterone to maintain endometrial receptivity if fertilization occurs. Without trophoblast implantation and decidualization, the corpus luteum regresses, causing a marked decline in circulating progesterone levels. This triggers a local inflammatory response in the endometrium involving leukocyte infiltration, cytokine release, edema, and activation of matrix metalloproteinases (MMP-1 and MMP-3) and strong contractions caused by prostaglandins. The result is tissue rupture and the fall of the upper two-thirds of the endometrium, the functional layer, during the menstrual phase of the cycle. We conducted a study on the origin of the main proteins that act in endometrial degradation, we can draw an evolutionary line to unravel the mystery surrounding this event. Using text mining, the MedlineRanker and PESCADOR programs, we have constructed a pathway for endometrial degradation comprising 28 genes. The protein sequences were used to build FastTree phylogenetic trees after multiple alignments with Muscle, and the ladistic distribution was analyzed with the TaxOnTree web tool of our group. The origin of the pathway requires the appearance of placentals, with some parts appearing in mammals and one is restricted to the order Catarrhini. For example, the LEFTY-2 gene, an inhibitor that positively regulates the transcription of ECM-degrading MMPs, originated in Catarrhini, however, the other member of his family that acts on the pathway, LEFTY-1, appeared already in Eutheria. According to this scenario, Plasmines, which also participate in ECM's degradation process, emerged in Theria. However, Plasminogen PLAU, the plasmin precursor, can only be converted by the PLAUR receptor when it emerged in Eutheria and its ligand, which is plasminogen itself, originated in Mammalia. In conclusion, the pathway has a recent origin bias, however the data we present do not map functions restricted to the family Hominidae of the great primates and more studies are needed.

Funding: CAPES Computational Biology Networks: Biologia Sistêmica do Câncer, BSC.

””

Classification, diversity and structural analysis of Ammonium Transporters

Gilberto Hideo Kaihami, Aureliano Coelho Proença Guedes, Gabriel Sánchez Hueck, Gianlucca Gonçalves Nicastro, Robson Francisco de Souza

Abstract

Nitrogen is an essential building block for many biological macromolecules. Its transportation through the cell membrane occurs in the protonated state (NH4+) or in the reduced form (NH3) and is mediated by proteins belonging to the AMT/MEP/Rh superfamily. Although the importance of this transporter family has long been recognized, the evolutionary relationships of several experimentally characterized homologs and the transport mechanism remain obscure. The current literature recognizes most members of the AMT and Rh families from eukaryotes, while homologs of MEP are found across all domains of life. To better understand the distribution and evolution of this superfamily, we systematically searched for homologs in the NCBI's protein database and inferred phylogenies for both the superfamily and each of its member families. Our results confidently assign to the AMT family several bacterial and archaeal homologs and revealed a well defined monophyletic group, comprising several instances of fusions of AMT homologs to signal transduction output domains, such as kinases, nucleotidyl cyclases, chemotaxis proteins and DNA binding domains. These fusions are typical of gram-negative bacteria. Gram-positive bacteria and Archaea often encode fusions to the regulatory protein P-II that belong to two other, unrelated, clades. Recent experimental work has shown that an AMT domain fused to a histidine kinase, isolated from a gram-negative bacteria, lost its transporter activity and acts as a sensor for extracellular levels of ammonium. Analysis of the structure of the histidine kinase's sensor domain suggests the signal transduction fusion, and the associated functional shift, may have been favored by the localization of the NH4+ ion in AMT-like transporters, where it occupies a non-canonical pore and is stabilized by residues distinct from those observed in the MEP and Rh proteins. Based on these results, we suggest a novel mechanism for transport and sensing of ammonium for the members of the AMT family and highlight experimental evidence that supports the hypothesis that all members of the AMT family fused to signal transduction domains lack its transport activity and function only as a sensor.

Funding: Capes, Fapesp

”

Classification of Substrate Binding Proteins in a Signal Transduction Context

Aureliano Coelho Proença Guedes, Gilberto Hideo Kaihami, Robson Francisco de Souza

Abstract

Sensing environmental changes and relaying this information to inside the cell is essential for all organisms. Transmembrane signal transduction proteins possess one or more extracellular sensory domains and cytoplasmic domains connected by membrane-spanning regions. Changes in environmental conditions or the presence of external chemical stimuli are detected by the extracellular domains, leading to structural modifications in the cytoplasmic portion and modifications of intracellular targets that affect cellular behavior, such as protein post-translational modifications or synthesis of small molecules that act as secondary messengers. The well-known substrate-binding proteins (SBP) correspond to a set of three distinct folds that contain two symmetrical internal subdomains, linked by a hinge region which promotes a "venus flytrap"-like motion. These domains have solute binding activity and may work either as components of transport or of signal transduction systems. Recent classifications of SBPs have identified families associated with these two functional categories, but whether transitions between these categories or episodes of recruitment to new functions have not been systematically explored. Additionally, recent work has demonstrated physical interactions between SBP proteins and periplasmic CACHE domains of transmembrane signal transduction proteins, suggesting that the evolution of signal transduction-associated SBPs, at least in some cases, may have originated after the emergence of interactions with the Cache domains. In this work, we collect SBPI, SBPII and SBPIII proteins from a set of prokaryotic genomes and performed sequence similarity clustering to identify SBP families associated with distinct functions. Our goal is to verify whether the association of a single-family to different functions could reveal episodes of recruitment of SBPs to novel functions. Our results show that many SBP families participate in transport and signal transduction systems, suggesting multiple independent functional shifts. Also, SBPs in signal transduction systems are more often directly fused to transmembrane regions bound to cytoplasmic effector domains instead of associated with Cache domains. This suggests that Cache-independent evolutionary pathways, such as direct recombination between SBPs and effector proteins could be a source of diversity in this system. We are now working on a detailed functional annotation of the SBPs and Cache in an effort to understand whether Cache-associated SBPs are biased to sensing the concentration of a limited set of solutes.

Funding: CAPES, FAPESP

”

The origin of the stoma was during the Paleozoic era, as ancient genes were co-opted by the stoma system

Beatriz Moura Kfoury de Castro, Tetsu Sakamoto, José Miguel Ortega

Universidade Federal de Minas Gerais

Abstract

Stomata are typically found in plant leaves but can also be found in some stems. These structures comprise specialized cells known as guard cells surrounding stoma that function to open and close the stomatal pores, allowing plant to take in carbon dioxide, which is needed for photosynthesis. They also help to reduce water loss by closing the structure when conditions are hot or dry. In Arabidopsis, cell lineage undergoes a series of cell divisions and cell-state transitions to produce a stoma. An evolutionarily ancient group that are known as basic helix-loop-helix (bHLH) transcription factors, specify cellular identity in both plants and animals. These transcription factors regulate the transition and differentiation events through the lineage of plants. Some genes associated with regulating stomatal differentiation are also associated with hormonal and environmental stress responses. To retrieve genes associated with stomata, scientific papers were collected describing stomatal development in Arabidopsis thaliana. These genes were analyzed and selected manually. After this selection, a regulatory pathway was constructed with manual curation. To analyze the evolutionary origin of genes in the pathway, their sequences were recovered from the UniProtKB database and their orthologs from other species were recovered using TaxOnTree software produced by our group. To determine the lowest common ancestor (LCA) between species with representative proteins, we used the cluster of orthologs for inferring the clade of origin of the gene. As a result, we were able to characterize the origin of 52 genes linked to the stoma system in Arabidopsis thaliana. Inference from gene origin showed that most of the genes are related to cell division and hormone and enviromental signaling and they appeared in the more ancient clades of the evolutionary history of plants, approximately during the early Mesoproterozoic (1800-1300 Ma), in the Eukaryota clade. Notably, most of the genes had originated during early Paleozoic era (540-480 Ma) between a middle Cambrian – Early Ordovocian interval, in the Embriophyta clade. These genes are mainly linked to the differentiation of stomatal tissue. Other genes originated more recently in plant evolution, such as Magnoliphyta and Mesangiospermae clades, during the middle Jurassic (175 Ma). Our data showed that the main genes involved in controlling stoma formation probably originated in Embryophyta, during the conquest of the terrestrial environment by plants. The stoma structure controls the gas exchange by the plant and it is directly related to essential processes for plant survival, such as respiration, transpiration and photosynthesis. Although several genes originated more anciently in the early Mesoproterozoic, these genes were probably later co-opted to form and control the stoma system, as they linked an environmental and hormonal response role, since during this period there was a large accumulation of carbon dioxide in atmosphere. Linking the networks that control stomatal development might improve our understanding of evolutionary history of plants and exemplifies how the response to the environment is important for understanding the origin of some current structures and their functions.

Funding: CAPES Computational Biology Networks: Biologia Sistêmica do Câncer, BSC.

,,,,,,,

Characterization of the mitochondrial genome of Phellinotus piptadeniae (Basidiomycota, Hymenochaetales) and insights on the phylogeny of Agaricomycetes through comparative mitogenomics

Daniel Silva Araújo, Paula Luize Camargos Fonseca, Gabriel Quintanilha Peixoto, Ruth Barros, Bertram Brenig, Vasco A de C Azevedo, Elisandro Ricardo Drechsler dos Santos, Aristóteles Góes Neto

University Göttingen

Abstract

The Fungi kingdom includes extremely diverse organisms and its taxonomy is constantly going through changes because of its diversity and discovery of new species. The elucidation of the phylogenetic relationships between fungi by analysis of mitogenomes is a common method, mainly because the mitochondrial genome is well conserved between different groups of fungi. In this work, the mitogenome of the wood-decaying fungus Phellinotus piptadeniae, a new species and genus of Agaricomycetes described by our group, was sequenced and assembled using the programs FastQC and SPAdes. The mitochondrial contig was identified with a BLASTn search against other mitogenomes, and the annotation was done using MITOS2, MFannot and RNAweasel. Next, we used the available Agaricomycetes mitogenomes in NCBI and JGI to build a phylogenetic tree including P. piptadeniae. Available fungal mitogenomes were selected and annotated using aforementioned programs and an alignment was built using MAFFT program based on the 14 core mitochondrial genes, rps3 protein and two ribosomal RNA. The mitogenome of P. piptadeniae has a size of 137, 790 bp with 24.1% GC content, 14 mitochondrial core-genes, 2 rRNAs, 30 tRNAs, 47 introns and 62 unidentified open reading frames. We also found coding domains for homing-endonucleases (HEs) (GIY or LAGLIDAGD) and dpo. For our comparative mitogenomics analysis, a total of 55 mitogenomes were retrieved from public databases and different genes, whose locations in the genome were conserved in the species of a same family or genus, were identified, such as: (i) atp6, cob, cox1, cox2, cox3, nad2, nad4, nad4l, nad5, rrnL and rrnS for Ganoderma genus; (ii) atp6, atp9, cob, cox1, cox2, cox3, nad1, nad2, nad3, nad4l, nad5 and nad6 for Hymenochaetaceae family; and (iii) also all the 17 genes for Russulaceae family. Deviations of these patterns may be explained by the presence of HEs in these mitogenomes: from all the mitogenomes used in our study, there were coding domains for HE in 50 of them. The phylogenetic tree was built by distance-based method on the genetic set comprised of the 14 core-genes, rrnL, rrnS and rps3. The use of the mitogenome as a tool for a better understanding of fungal phylogenetic relationships proved to be useful but the deficiency of enough mitogenomes of all different families in the Agaricomycetes class in public databases still is a problem to overcome since unbalanced sampling influences on the quality of the results.

Funding: CNPq

’’’’’’’’’’’’’

GENOMIC SURVEILLANCE OF ZIKA AND CHIKUNGUNYA VIRUS IN MINAS GERAIS, BRAZIL

Felipe Campos de Melo Iani, Marta Giovanetti, Jaqueline Goes de Jesus, Talita Émile Ribeiro Adelino, Maira Alves Pereira, Joilson Xavier dos Santos Junior, Vagner de Souza Fonseca, Julien Theze, Ester Cerdeira Sabino, Marluce Aparecida Assunção Oliveira, Aristeu Mascarenhas da Fonseca, Flavia Salles, Nuno Rodrigues Faria, Luiz Carlos Junior Alcantara

Laboratório de Genética Celular e Molecular, ICB, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil

Abstract

Recent emergencies of chikungunya (CHIKV) and zika (ZIKV) viruses have raised serious concerns due to the virus' rapid dissemination into new geographic areas and the clinical features associated with infection. Using a combination of portable whole genome sequencing and molecular clock analyses, we investigated the genomic diversity and epidemiological dynamics of CHIKV and ZIKV in different municipalities of Minas Gerais state (MG), Southeast Brazil. MG had the first CHIKV confirmed case in September 2014 but was confirmed as an imported case from Venezuela. MG has since 2014 up to date, a total of 31, 095 probable cases and of these, 275 CHIKV cases have been RT-qPCR laboratory-confirmed. The first Brazilian cases of autochthonous transmission of the ZIKV were confirmed in May 2015 in northeast Brazil. However, the first laboratory confirmed case in MG was in February 2016. MG has since 2014 up to date, a total of 15, 644 probable cases and of these, 1, 609 ZIKV cases have been RT-qPCR laboratory-confirmed. We generate 14 CHIKV and 7 ZIKV near-complete genomes sequences of virus isolate obtained directly from clinical samples. Our phylogenetic reconstructions indicated the co-circulation of the CHIKV - East-Central-South-African (ECSA) lineage and the ZIKV Asian genotype in MG state. Time-measured phylogenetic analysis revealed the CHIKV ECSA lineage was introduced in Minas Gerais around November 2014 (95% Bayesian credible interval: May 2014 to April 2015) and that the ZIKV – Asian genotype was introduced around July 2014 (95% Bayesian credible interval: May to December 2014). Beside CHIKV circulation in the state has been quickly detected by the surveillance system, our data indicate that ZIKV probably has circulated unnoticed for 18 months before the first confirmed laboratory case. These findings reinforce that continued genomic surveillance strategies are needed to assist in the monitoring and understanding of arbovirus epidemics, which might help to attenuate the public health impact of infectious diseases.

Funding: Fapemig, CNPq, Capes, ZIKAlliance, Royal Society and Wellcome Trust Sir Henry Dale Fellowship, John Fell Research Fund, ERC,

”

Evolution of lignocellulose degradation characterizes the adaptation for heterotrophy to carbohydrates that appeared in Fungi

Fenícia Brito, Tetsu Sakamoto, José Miguel Ortega

Universidade Federal de Minas Gerais

Abstract

Lignocellulose is the most abundant terrestrial biopolymer on Earth. Cellulose is a high molecular weight linear homopolymer of repeated units of glucose, containing highly crystalline regions and less-ordered amorphous regions. Hemicellulose is a linear and branched heterogeneous polymer typically composed of five sugars: L-arabinose, D-galactose, D-glucose, D-mannose and D-xylose, and acetic, glucuronic and ferulic acids. The backbone of hemicelluloses chains can be a homopolymer or a heteropolymer. Lignin is a very complex molecule constructed of phenylpropane units linked in a large three-dimensional structure. Lignin is closely bound to cellulose and hemicellulose and are particularly important in the formation of cell walls, especially in wood and bark, because they lend rigidity and do not rot easily. Fungal ability as decomposers of organic matter is widely known. When it comes to the degradation of lignocellulose fungi also stand out for their efficiency. Although some bacteria are known to decompose the lignocellulose polymer at some level, we wondered which organisms have the essencial lignocellulosic enzymes and hence when the lignocellulose degradation has emerged. In this work we investigated 7, 957 proteomes from the UniProt Reference Proteomes database for the set of enzymes for lignocellulose degradation. Using software TaxOnTree to performed phylogenetic and taxonomic distribution analysis, we were able to identify which organisms present which proteins and described their potential mechanism for decomposing the lignocellulose polymer. Some organisms can decompose the celulose fiber from its ends, by using a set of two enzymes: Cellobiose dehydrogenase or Cellobiohydrolase and Beta-glucosidase. Besides fungi, these enzymes were found in bacteria and in some metazoans. Taxonomic distribution suggests either a Lateral Gene Transfer (LGT) event or great deletions from an ancient ancestor whether the enzymes lack function. The two enzymes implicated in the breakdown of the xylan portion of hemicellulose, Beta-xylanase and Beta-xylosidase, were found in two orders of Halobacteria, different bacteria phyla (e.g. Bacteroidetes, Acidobacteria, Proteobacteria and Actinobacteria), in fungi and plants. The two enzymes for the breakdown of the manan backbone in hemicellulose, Beta-mannosidase and Endo-1, 4-beta-mannosidase, were found in fungi, plants and a few bacteria phyla (Bacteroidetes, Proteobacteria, Actinobacteria and Chloroflexi). As for the lignin decomposition, the main enzymes involved in these process, ligninases MnP, LP and VP in the Agaricomycetes class, although some other peroxidases, also found in bacteria, can also participate of the lignin breakdown. Remarkably, only fungi have the complete set of enzymes for decomposition of the entire polymer of celulose, hemicelulose and lignin. These results suggests that either LGT was a driving force on spreading these enzymes in species of bacteria and metazoans, or great event of deletions avoided the fixation of such activities originated in a very ancient ancestor. By inspecting the distribution, we favor the possibility of the effective digestion have been originated in fungi with ocasional LGT events. It seems that the

,

Comparative genomics of R-body determinants

Gabriel Sánchez Hueck, Robson Francisco de Souza

Abstract

Refractile (R) bodies are potential toxin-delivery proteic structures able to stretch and puncture cell membranes on a pH-dependent manner, whose presence have been verified in some free living and endosymbiont bacteria, but the role and genomic context of reb genes responsible for their polymerization is not yet understood. RebB domain genes appear on genomes of proteobacteria, bacteroidetes and acidobacteria, generally displaying several contiguous duplications, but operon configuration exhibits varied forms and distribution. R-bodies have been shown to require the presence of distinct reb family members, and sometimes even non-homologous genes, to ensure polymerization. Regulatory genes have been demonstrated, but their taxonomic distribution and domains have not been fully determined. The precise order of events guiding horizontal transfer and duplications in rebs and neighboring genes also remains undefined. We intend to explore the presence of reb homologs and identify new genes associated with R-body function, allowing the inference of their evolutionary history. Potential genes identified on our work will serve as candidates for experimental characterization in future essays, with special interest on of toxins and antitoxins. At the core of our strategy we make use of iterative sequence alignment tools such as jackhmmer and Psi-Blast for collection of homologs, domain recognition tools, such as hmmscan and hhsearch for classification and labeling, mmseqs as a tool for grouping sequences and reducing redundancy based on the e-value and identity of pairwise alignments, and FastTree and figtree for the construction and visualization of inferred phylogenies. We have also made use of tools for gene locus and neighborhood collection developed in our laboratory. Our main sources of data are the non-redundant RefSeq database for protein sequence collection and Pfam for models for domain description. We have found that reb family members have a conserved common core and variable regions on the N and C-terminal portions, which might account for the distinct polymerization events that occur during R-body assembly. We've created models based on the topology of the phylogenetic trees that the reb family displays, which has aided the visualization of reb loci and the distinct patterns displayed by them on genomes, and will be useful to describe events of horizontal transfer. We have additionally detected a high frequency of regulatory proteins neighboring rebs, and expect to identify, through the collection of homologs of known regulators as well as through mutation correlation measures, whether or not they might regulate the reb operons.

Funding: FAPESP - Iniciação Científica

,

Comparative genomics of the type four secretion system pumping ATPases VirD4/VirB4 from the FtsK–HerA superfamily reveals a new clade in the Candidate Phyla Radiation.

Gianlucca Gonçalves Nicastro, Robson Francisco de Souza

Abstract

Bacteria are unicellular organisms that participate in a diversity of biological interactions in the environment. One of the most important of these interactions is the transfer of genetic material between different organism. This mechanism is known as horizontal gene transfer and plays an important role in bacterial evolution and the emergence of antibiotic resistance. The type IV secretion systems (T4SS) are large multi-protein structures involved in the single-stranded DNA transfer. The components of this system can be divided into three main parts: the pump ATPases, the elements that form the translocation channel and the proteins that form the pilus. T4SS genes are widely spread among prokaryotes, including Archaea. Previous phylogenetic studies of the T4SS VirB4/D4 ATPases demonstrated a robust separation in eight large clades, that in part reflects the structure of the cell envelope and were used as the basis of a phylogenetic classification of T4SS. Despite its broad covering, such studies lack information about the most recent sequenced genomes, as the DPANN archaea and the bacterial candidate phyla radiation (CPR). In this study, we sought to expand the T4SS classification based on the VirD4/B4 phylogeny. In order to reliably identify all VirD4/B4 homologs, we performed iterative sequence similarity searches against NCBI's non-redundant database using representatives of the most closely related families within the FtsK–HerA superfamily, including members of the VirD4, VirB4, HerA and FtsK families. After removal of redundant sequences, careful phylogenetic analysis revealed that the HerA, VirD4 and VirB4 proteins form well defined monophyletic groups. The VirD4 and VirB4 subtrees were found to contain clades similar to those identified in previous works, but revealed the presence of a new group formed mostly by members of CPR. We also observed the presence of a new clade of VirD4 genes of Actinobacteria. Our classification thus reveals a new type of T4SS from CPR bacteria, that could be related to a singular cell envelope structure within this clade.

Funding: Fapesp, Capes

,

Phylogenetic analysis of the TRAFAC class and discovery of the first prokaryotic septin

Guilherme Bastos Gomes, Robson Francisco de Souza

USP

Abstract

Septins are GTPases capable of assembling into hetero/homo-oligomers that were first described as components of a ring structure at the bud neck formed during cell division in Saccharomyces cerevisiae cells. The name septins is due to the localization of this ring at the septal collar. Posterior studies have shown that septins interact with actin and microtubules, are able to form filaments and can regulate membrane dynamics, such as blebbing formation and flexibility. Given their role in membrane dynamics and its ability to form scaffolds, septins were dimmed the fourth element of the cellular cytoskeleton, together with actin, tubulin and intermediate filaments. Although initially identified only in animals and fungi, recent studies have shown a wider distribution of this protein family across many eukaryotic lineages, such as red and brown algae. Interestingly, although many related small gtpases are known from prokaryotes, no septin homolog was ever found in any bacterial genome. To understand the origins and distribution of septin-related proteins, we performed exhaustive searches for homologs of members of the TRAFAC (Translation factor-related GTPases) group, a wider assembly of small gtpases that includes septins. By focusing on a paraphyletic group known as paraseptins, we were able to recover all the closest family members of septins within the TRAFAC class, including Elongation Factors, OBG-like, RsgA, Era, TrmE, MnmE, Rab, Ras, Dynamins, GIMAPs, Tocs and Septins. Sequence similarity clustering revealed closely related groups, from which representative sequences were gathered and aligned for phylogenetic analysis. Several paralogs of the GIMAP family of paraseptins were found in groups of Teleostei, what agrees with the hypothesis of a fourth wide genome duplication in this lineage. The diversification of the GTPases of immunity associated proteins were also analysed and this protein is spread among non-opisthokont Eukaryotes. This analysis also revealed a close relative of the septin group spread among several bacterial clades and closely related to a clade of the Chlamydomonas reinhardtii septins. This bacterial putative septin is fused with a domain of unknown function composed of one soluble region surrounded by two transmembrane regions before a cytoplasmic region and an N-terminal region also associated with several other GTPases. The discovery of a close relative of septins spread across many bacterial lineages can shed new light on the evolution and origin of this important group of proteins and give material for further experimental analysis.

Funding: CAPES

”

TOLL receptor gene family evolution in insects

Letícia Ferreira Lima, Rodrigo Jardim, Renata Schama

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

Abstract

Arthropoda can be found in almost every habitat on earth and many species are intimately related to human life, for instance, arthropod-born diseases as malaria and yellow fever. Insects presence in most habitats and their wide variety of diet and behavior also means that they encounter various microorganisms many of which may be pathogenic, as result insects have evolved mechanisms for recognition and elimination pathogens, which include the signaling Toll pathway. Toll receptor is transmembrane protein essential for embryonic development and immunity, the induction of the Toll pathway by Gram-positive bacteria or fungi leads to the activation of cellular immunity as well as the systemic production of certain antimicrobial peptides. The Toll receptor is activated when the proteolytically cleaved ligand Spatzle binds to the receptor, eventually leading to the binding with other protein of pathway and activation of NF-kB factors. To date, nine genes have been found in Drosophila melanogaster genome and similar numbers were found in other insects. Insects are ideal models for the study of the diversity of gene families and their evolutionary mechanisms because they have a well studied and well-known phylogeny and there are many examples of evolutionary specialization that have arisen in different bloodlines, such as hematophagy. Phylogenetic analyses using the Toll domain of each sequence retrieved from 40 insects genome was able to divide the Toll family into three well supports clades. The results revealed that insect Toll domain formed three major clusters. Here, we have shown that there is a variety of Toll copies in insect groups, with duplications and gene losses, our findings also show that the Toll9 does indeed appear to be closer to vertebrates than the other groups, and may indicate that it may be the most closely related group. In this study, Toll9 genes in the Hymenoptera order were not found, which may suggest that the gene was lost in the order. Insects immune evolution like the Toll pathway is important for an understanding of vector biology and behavior, helping in aspects of vector control and disease transmission.

Funding:

,,,

Profile HMMs as auxiliary tools for the taxonomic classification of viruses: a case study using Spounarivinae phages

Liliane Santana Oliveira Kashiwabara, Miriã Nunes Guimarães, Wendel Hime Lima Castro, Arthur Gruber

USP

Abstract

Profile HMMs are probabilistic models that are much more sensitive to detect remote orthologs than conventional pairwise alignment methods. We have recently developed TABAJARA, a program for the rational design of profile HMMs. In this work, we report the development and application of profile HMMs for the detection and taxonomic classification of phages of the subfamily Spounavirinae. We obtained a dataset of bona fide Spounavirinae sequences from the NCBI's Identical Protein Groups (IPG) database, restricting the query to complete sequences of terminase and tail sheath protein (TSP) associated to txid857473. Protein sequences were aligned with MUSCLE and the resulting alignments processed by TABAJARA to produce profile HMMs specific at the levels of genus and subfamily. Specificity and sensitivity of all models were assessed by similarity searches against the training set. As a final validation procedure, we tested all models against a dataset of Spounavirinae-depleted Myoviridae sequences. These tests revealed the detection of 60 unique sequences of terminase and 66 of TSP, corresponding to 72 non-redundant viral genomes. To clarify whether these sequences represented false positives detected by the models or corresponded to misclassified sequences, we inspected their taxonomic assignment on the IPG and NCBI Taxonomy databases. With no exception, all sequences belonged to orphan genera (not included within any subfamily) or to unclassified viruses. Terminase and TSP sequences of this group, together with representatives of all Myoviridae subfamilies, including the Spounavirinae subfamily, were used in ML phylogenetic reconstructions with FastTree program. Trees derived from both protein datasets revealed that all sequences identified by our models constituted a monophyletic group, including sequences originally classified as Spounavirinae, as well as sequences from the orphan genera Cp51virus, B4virus, Bastillevirus, Wphvirus, Bc431virus, Nit1virus, Agatevirus and Sep1virus, and some previously unclassified Myoviridae viruses. This wide group of taxa corresponds to Herelleviridae, a new virus family recently proposed by Barylski et al. (Syst Biol. 2019 May 25. pii: syz036) using a variety of phylogenetic analyses based on genomic, proteomic and marker gene-based data. Similar results were also observed by Aiewsakun et al. (J. Gen. Virol. 99: 1331-1343, 2018), using a genetics-based platform that computes sequence relatedness between viruses. Altogether, this body of evidence suggests that our models, rather than detecting false-positives, had indeed identified mis- or unclassified Myoviridae sequences. We conclude that profile HMMs may be used as auxiliary tools for the taxonomic classification of known and emergent viruses.

Funding: CAPES

,,

Functional genomics of the Rhipicephalus microplus tick infection process by Metarhizium anisopliae: unraveling the mechanisms of host-pathogen infection

Mateus Martins Frasnelli, Ana Trindade Wink, Claudia Elizabeth Thompson

UFCSPA

Abstract

Metarhizium anisopliae is an entomopathogenic fungus that causes infections in several arthropod species. The biological control of parasitic arthropods such as the tick Rhipicephalus microplus has great economic and sanitary interest, since they bring losses in Brazilian livestock. Several acaricides are available in the market, but they leave residues in meat, milk and derivatives. Sustainable alternatives have been studied and one of the best-known models is Metarhizium anisopliae. However, the development of an efficient control method is hard because it is yet not possible to completely understand their infection mechanisms. This project aims to help in the understanding of the evolutionary history of these organisms and pathogenicity related genes. The research was performed considering the pathogenicity, size, assembly status, GC content of the Metarhizium genomes deposited in public databases. We identified ortholog gene groups in the 12 species found using OrthoFinder. These sequences were aligned with PRANK software and, subsequently, a supermatrix was constructed from multiple alignments with SCaFos software. Finally, distance and probabilistic methods of phylogenetic reconstruction were applied with the MEGA software, and these processes were documented and used in the building of a pipeline. We selected 5, 509 groups of orthologous genes, from which we obtained a phylogenomic tree with high statistical support. A selected tree depicts an evolutionary relationship between the twelve genomes of the genus Metarhizium, corroborating some of the data available in the literature. We were able to identify genes shared among different species and those specific to each organism. This has allowed the establishment of the relationships among Metarhizium species and the identification of genes related to pathogenicity.

Funding: FAPERGS

”

SWeeP and machine learning in supertree construction: family Formicidae analysis

Monique Schreiner, Roberto Tadeu Raittz, Mariane Golçalves Kulik

Universidade Federal do Paraná

Abstract

Phylogenies including all taxa of a large group, the supertrees, are essential for research in macroevolution, biogeography and conservation. In groups with lack of data, the construction of super trees becomes complex. Particularly, the ants (family Formicidae) compose a group with genetic data distribution quite heterogenous wherein few species have thousands of sequences registered on biological databases while most of the species have few or none sequences registered, fact that make a global phylogenetic analysis difficult. The present research proposes the construction of a supertree for all extant ants, including the ones missing genetic data available, using vectorial techniques and artificial intelligence. To perform this, all ant protein sequences available on NCBI were downloaded and clustered. For prospection, the largest cluster were analyzed. The sequences of the largest cluster were vectorized using the algorithm SWeeP which transforms sequences of amino acids in compact vectors preserving comparability of the sequences and allowing big data analysis. One matrix per protein was generated for the most frequent proteins on the cluster. A Principal Component Analysis (PCA) was performed on the SWeeP matrixes. The result showed that it was possible to distinguish the main subfamilies of ants using a few of the principal components. Multilayer Perceptron Neural Networks (MLP) were trained to classify the subfamilies using the first one hundred principal components of the matrixes. The mean accuracy of the neural networks was 0.9576 (SD=0.064). The neural network with the best performance was for the protein Cytochrome C Subunit I with 0.994 accuracy. The preliminary results showed that it is possible to classify organisms using molecular data without the use of alignment techniques. Alignment techniques are computationally expensive and hence limit the amount of data that can be used. The classification learning model corroborates the potential of the proposed vectoral model. The learning model will be applied in the prediction of missing elements and will allow the use of multiple proteins. At last, the developed model will be used in the construction of the Formicidae supertree.

Funding: CAPES

,,,

Reconstructing the phylogeny of Corynebacteriales while accounting for Horizontal Gene Transfer

Nilson Coimbra, Aristóteles Góes Neto, Vasco A de C Azevedo, Aida Ouangraoua

Université de Sherbrooke

Abstract

Horizontal Gene Transfer (HGT) is a common mechanism in bacteria that affects the genomic content of extant organisms. However, most traditional methods for bacterial phylogeny reconstruction assume only vertical inheritance in the evolution of homologous genes. Here, we present a new method for bacterial phylogeny reconstruction that accounts for the presence of genes acquired by HGT in genomes. A gene tree-based method was devised to identify and correct putative transferred genes (PTG). The method is applied to the reconstruction of the phylogeny of the Order Corynebacteriales, the largest clade in the Phylum Actinobacteria.

Funding:

,,,,,,,,,,

Introns and homing endonucleases shape mitochondrial genomes of fungal species from Hypocreales order (Ascomycota)

Paula Luize Camargos Fonseca, Ruth Barros, Daniel Silva Araújo, Dener Eduardo Bortolini, Gabriel Quintanilha Peixoto, Vasco A de C Azevedo, Bertram Brenig, Luiz Eduardo Vieira Del Bem, Fernanda Badotti, Aristóteles Góes Neto, Eric Roberto Guimarães Rocha Aguiar

University Göttingen

Abstract

The order Hypocreales is composed of ubiquitous and ecologically diverse fungi classified in saprobes, biotrophs and pathogens of other species. One of the main genera is Trichoderma, which are used as biocontrol and biofertilizers agents for plant growth. Due to the variety of ecological functions presented by species from Hypocreales order, comparative genomics is an important tool to understand the differences observed in the fitness of these organisms. The mitochondrial genome (mtDNA) play an important role, providing energy to the cells and regulating processes related to immune response. However, although its importance, the mechanisms that shape fungal mtDNA still poorly understood. To better understand mechanisms involved in the variability and evolution of mitochondrial genomes we investigated fungal species from the Hypocreales order. First, we sequenced and annotated T. harzianum mitochondrial genome, which was compared to others 34 mtDNAs species that were publicly available. Comparative analysis revealed the considerable elasticity mtDNAs, with length ranging from 24, 565 to 103, 844 pb. Although the size variation observed in mitochondrial genomes, gene copy number, size and structure of coding elements were highly conserved, suggesting that differences is likely on non-coding regions. Among the elements classified as non-coding regions, introns and homing endonucleases genes (HEGs) were the main contributors to the size variations. 267 out of 332 identified introns showed sequence similarity between species. The most fragmented genes (rrnL and cox1) exhibited the highest frequency of HEGs within intronic regions. In the genes with the lowest frequency of fragmentation (rrns and atp8), HEGs and introns were absent. We also investigated the possible transference of mitochondrial genes to the nuclear genome (NUMT). The gene nad5 was the most widespread in the nuclear genomes. In contrast, the genes atp8, atp9 and cox3, events of transference were not identified. Since the genes atp8, atp9 and cox3 are unique to all mtDNA evaluated, they were used to construct a time-scaled phylogenetic tree to estimate the origin of the order based on mitochondria information and to determine whether the presence of fragmented genes, introns, HEGs and NUMTs were related to time divergence. However, a weak association was found, indicating that other mechanisms could be responsible for the abundance of introns and HEGs. Altogether, our results indicate that HEGs and introns play an important role on the shaping of mitochondrial genomes, whether on fragmentation, duplication or transference of genes to the nuclear genome.

Funding: Conselho Nacional de Desenvolvimento Científico e Tecnológico

,,,,,,,,

Mitogenome data reveals strong differentiation among the isolated populations of Heliconius hermathena: a white sand ecosystem specialist.

Pedro de Gusmão Ribeiro, Renato Rogner Ramos, Darli Massardo, Marilia Lion, Marcio Zikan Cardoso, Marcus Kronforst, André Victor Lucci Freitas, Marcelo Mendes Brandão, Karina Lucas da Silva Brandão

UNICAMP

Abstract

Cycles of forest retraction and expansion during the Pleistocene presumably played a crucial role in the diversification of neotropical species by the formation of isolated forest refugia. Similarly, these cycles generated the strongly isolated pattern of the Amazonian white sand ecosystems: non-forest habitats with white sandy soils surrounded by forest matrix. As previously inferred by ecological and morphological data, such processes may have led to the isolation and diversification of subpopulations of H. hermathena, a butterfly endemic to these ecosystems with seven subspecies identified by their color patterns. Nevertheless, the genetic differentiation and structure among H. hermathena subpopulations are still unknown. We sequenced the mitogenomes of 71 individuals across six of the seven subespecies of H. hermathena from eitgh different localities. We then performed bayesian phylogenetic inference and population structure analyses in order to analyze patterns of differentiation among H. hermathena subpopulations and their phylogenetic relationships. Most of the analysis were performed in user-friendly platforms such as Galaxy Project and Geneoius 10, which implements most of the state-of-the-art bioinformatics programs. Currently, we are accessing the divergence times among these subpopulations to infer the specific mechanisms regarding the group's phylogeography and evolution. We show that two populations with equal wing color pattern (H. h. sheppardi) sampled from two different localities exhibit high genetic divergence and population structure. Conversely, a pair of phenotipically divergent subspecies (H. h. vereatta and H. h. duckeii) from two near sample sites in Faro, are genetically similar and have lower fixation index when compared to other sample localities. Furthermore, we found highly distinct haplogroups among H. hermathena subpopulations, with each haplogroup strongly structured in its own sampling locality. Our results suggest that the fragmented pattern of the white sand ecosystems may have actually played an important role in the formation and maintenance of differentiated populations and subspecies of H. hermathena. For this butterfly, the forest function as a barrier for free gene flow among its current populations.

Funding:

”

MITGARD: an automated pipeline for mitochondrial genome assembly based on RNA-seq data

Pedro Gabriel Nachtigall, Felipe Gobbi Grazziotin, Inácio L.M. Junqueira-de-Azevedo

Laboratório Especial de Toxinologia Aplicada (LETA), Instituto Butantan, São Paulo, Brazil

Abstract

In Metazoa, mitochondrial genes are the most commonly used markers for molecular species determination and phylogenetic studies due to their extremely low rate of recombination, maternal inheritance, ease of use and fast substitution rate in comparison to nuclear DNA. In this sense, the assembly of the mitochondrial genome is an important step to proceed with rapid species identification in biodiversity surveys as also a key tool to identify hidden lineages or cryptic species. Over the past decade, the emerging field of next-generation sequencing (NGS) has seen dramatic advances in methods and a decrease in costs. Consequently, we noticed a big expansion on data being generated by NGS, most of them from RNA-seq experiments aiming at different objectives. Since mitochondrial genes are expressed at different levels in the majority of animal tissues, mRNA sequences are usually co-sequenced within the target transcriptome, generating a sequence data that is commonly underused or discarded. Then, the design of a computational pipeline that can be easily and automated applicable to assembly mitochondrial genomes from RNA-seq data is a valuable tool in the constant expansion of high-throughput data generation. Here, we present MITGARD, an automated pipeline that reliably recovers and assembles the mitochondrial genome from RNA-seq data from various sources. MITGARD was developed using Python and third-party tools, by taking the RNA-seq data as input and confident mitochondrial genome assembly as output. The preliminary results, using RNA-seq data from venom glands of several species from the snake genus Bothrops, showed that MITGARD reliably assembled the mitochondrial genomes, which could be used in alignments and construction of phylogenetic trees. Our assemblies together with available mitochondrial genomes resulted in confident phylogenetic inferences in snake species. In this sense, MITGARD is a helpful approach to studies focusing to assemble and annotate the mitochondrial genome and/or sequences of specific mitochondrial genes that can be used for species identification and evolutionary studies.

Funding: This study was financed by FAPESP (Processes Numbers: 2016/50127-5; and 2018/26520-4)

”

Identification of intragenic retrocopies in chimaric transcripts in humans

Rafael Luiz Vieira Mercuri, Helena Beatriz da Conceicao, Pedro Alexandre Favoretto Galante

Instituto de Ensino e Pesquisa, Hospital Sírio-Libanês

Abstract

The evolution of a species occurs by the emergence of new genes. These can be originated through LINE1 (L1) mediated gene duplication, a phenomenon often encountered in the human genome. Although sometimes these genes are not functional, recent studies have shown that some retrocopies are transcribed and functional. Among all the retrocopies genes encoding a genome, those inserted into the intronic regions of (other) coding genes deserve attention. Because they are in a gene region, they may influence the transcription and post-transcriptional processing of the "host gene." It is currently known that there are retrocopies present in the human genome that are located in introns of coding genes. However, very little is known about the influence and contribution of these retrocopies in relation to their host genes. The aim of this work was to elaborate a systematic study of the retrocopies inserted in introns and exons of human coding genes (intragenic). The RCPedia (https://www.bioinfo.mochsl.org.br/rcpedia/) was used as a database to identify intragenic retrocopies in humans. Subsequently a comparison was made with GTEx (https://gtexportal.org/) to verify the expression profile in 53 human tissues of host genes and their intragenic retrocopies. First, we found 2499 intragenic retrocopies (990 in the same transcription strand and 1509 in the opposite strand to their host genes) and of these, 65% (1630 retrocopies) had their expression confirmed by GTEx. Testis presented the highest number of expressed retrocopies and bladder the lowest. Interestingly, , some of expressed retrocopies are located into genes involved in pathological pathways as Ras Homolog Family G (RHOG). Thus, our results bring important knowledge and can contribute to a better understanding of the origin of new genes and genetic novelties.

Funding: FAPESP

”

SPLACE: a tool to SPLit, Align and ConcatenatE genes for phylogenetic inference

Renato Renison Moreira Oliveira, Santelmo Vasconcelos, Guilherme Oliveira

Universidade Federal de Minas Gerais

Abstract

The production of phylogenetic trees containing multiple genes is best accomplished by concatenating all aligned genes into a supermatrix instead of generating a tree for each gene and then inferring the phylogeny by the consensus of all trees, i.e. a supertree. The advent of NGS technologies made it easier and cheaper to obtain multiple gene information from a large number of organisms of interest, generating a more robust supermatrix. The supermatrix, then, can be used in the phylogenetic reconstructions to generate a species tree. Many studies have used the supermatrix strategy to infer the phylogeny among species, such as when analyzing genomes from prokaryotic organisms and from eukaryotic organelle genomes (mitogenomes and plastomes). Building a supermatrix can be very time-consuming, especially if there is a large number of genes from many organisms to use in the analysis. Some published tools, such as SequenceMatrix, TaxMan, ScaFoS, TNT, and Phyutility, aim to concatenate gene files, but they require aligned gene files, what can be a problem with a large number of genes, as we mentioned. Here we present SPLACE, a tool to SPLit, Align and ConcatenatE the genes from all the species of interest to generate a supermatrix file, and consequently, a phylogenetic tree. To generate the supermatrix of n organisms, SPLACE will need n fasta files, each one containing all the g genes from a particular organism. First, SPLACE splits the genes from an organism, gathering the genes that have the same name from the n organisms into a single fasta file, therefore generating g new fasta files, each one containing the same gene from different organisms. Then, SPLACE aligns each one of the g fasta files using the MAFFT aligner, generating g' new fasta files. Finally, the genes in the g' fasta files that came from the same organism are concatenated into a single sequence, generating a single fasta file with the supermatrix containing n sequences, each representing one of the n organisms. Phylogeny is then reconstructed using the supermatrix fasta file. We used SPLACE to build a phylogenetic tree for all plant species with complete nuclear genome deposited on NCBI, using its respective chloroplast genes. The supermatrix was then submitted to CIPRES portal to generate a maximum likelihood phylogenetic tree using RAxML, with a bootstrap of 1000 replicates. The resulting phylogenetic tree showed the proper proximity among the plant species that belong to the same family.

Funding: 372439/2019-5, CNPq

”

Fastly evolving genes in parrots (Aves, Psittacidae) are associated with developmental processes

Thieres Tayroni Martins da Silva, Anderson Vieira Chaves, Francisco Pereira Lobo

Universidade Federal de Minas Gerais

Abstract

During the course of evolution of protein-coding genes, most non-synonymous mutations (dN) are removed from gene pools due to negative selection, leaving a footprint of synonymous mutations (dS). However, some homologous genes have an excess of dN substitutions when compared to dS, indicating a selective advantage for sequence variation over conservation. Such genes are referred as adaptive genes and have been demonstrated to be involved in major adaptive processes in vertebrates, such as reproduction and immunity. Therefore, the detection of genes evolving under positive Darwinian evolution is a prevailing strategy in comparative genomics studies to identify genes potentially involved in adaptation processes. Birds are subject to unique selective pressures due their diverse lifestyles, and previous studies found the genes with evidence of positive selection in this taxon to be associated with developmental processes, such as spinal cord and bone development. Among birds, Psittaciformes (parrots and relatives) are known by their unusual longevity and cognitive capabilities. In this work, we investigated candidate genes for traits relevant to Psittaciformes ecological and functional diversity using POTION2, a software developed by our group to search for genes evolving under evidence of positive selection. We used the complete genomes from 16 avian species representing major extant clades to search for fastly-evolving genes in Psittaciformes when compared to other Psittacopasserria. In a group of 16 high-quality complete genomes (10 Passeriformes, 4 Psittaciformes, 2 Falconiformes birds with BUSCO completeness greater than 0.89). We used MUSCLE for protein alignment, trimAL to remove poorly aligned columns (> 50% gaps), newick utilities for newick tree file manipulation and FastCodeML to infer positive selection. POTION2 requires a phylogenetic species tree to be executed in branch mode, which was obtained from TimeTree of Life website, and ML phylogenetic reconstructions conducted by us. Preliminary analysis of 7036 high-quality 1-1 orthologs found 230 genes (3%) with evidence of positive selection in Psittaciformes. In addition to the processes described in previous studies, we have also found genes related to the regulation of reproductive processes, embryological and cellular processes of development, stimulation and response to beta growth transforming factor, and oxygen depletion. Besides that, many of the positively targeted genes still represent uncharacterized proteins, and comprise interesting targets for functional characterization. Thus, these genes could have been important in the evolution of morphological, physiological and behavioral adaptive traits peculiar to each bird order.

Funding: CNPq

6 — Proteins and Proteomics

”

Proteomic approach for the evaluation of oxide nitric dependent pathways during Leishmania major infection

Adriene Yumi Ishimoto, Luiza A. Castro-Jorge, Dario Simões Zamboni

Universidade de São Paulo

Abstract

Leishmaniasis is a tropical and subtropical endemic disease caused by parasites of the genus Leishmania. The disease clinical manifestations depend on the infecting Leishmania species and the host immune response, and can be classified into two types: tegumentary and visceral. Macrophages represent the primary line of defense against infection, and nitric oxide (NO) production are one of the major mechanisms involved in eliminating parasites. In order to identify proteins involved in the parasite control through the nitric oxide pathway, we performed a proteomic analysis. Bone marrow derived macrophages (BMDMs) from wild type and NOS2-deficient C57BL/6J mice were infected with Leishmania major, and protein levels were quantified by mass spectrometry. The analysis was based on statistical calculations available in packages and functions of the R language, such as t.test function, and limma and ROTS packages. The differentially expressed proteins (DEPs) were defined as those that obtained p-value less than 0.05. Gsr, Arg1, F13a1, Pcna, Plin3 and Cd36 proteins were more differentially expressed in the context of Leishmania infection. Through pathway enrichment analysis using packages such as clusterProfiler, ReactomePA, WebGestalt and EGSEA, we identified activated pathways related to the regulation of adaptive immune response, immune effector process and leukocyte mediated immunity. Cd14, Cd36, Arg1, Ctsl, Dctn2, Rab7 e Arf1 were selected as acting in the regulation of the identified pathways by protein interaction analysis. The reliability of the detection of differentially expressed proteins was increased when using different approaches to statistical analysis, mainly when compared to traditional methods. The evaluation of the protein interaction network allowed to identify important proteins of modulated pathways during the infection, improving the understanding of infection control in the absence of nitric oxide.

Funding: Universidade de São Paulo

,,,

PRIORITIZING PROMISING COMPOUNDS IN VIRTUAL SCREENING CAMPAIGNS

Alexandre Victor Fassio, Rafaela Ferreira, Michael Keiser, Raquel Melo Minardi

Universidade Federal de Minas Gerais, UFMG

Abstract

Proteins are crucial macromolecules to all organisms as a whole, and countless diseases are associated with the proper functioning of proteins. Not surprisingly, proteins are the focus of numberless biological research whose focus is the development of new drugs able to modulate these macromolecules. However, the development and discovery of new lead compounds is a highly expensive and time-consuming program that takes up to 10-15 years. Thus, computational techniques like structural-based virtual screening (SBVS) and molecular docking contribute significantly to the early-stage drug discovery. A typical SBVS campaign consists of three major phases, namely the data preparation, the docking, and post-analysis. Commonly, a researcher starts with more than 20, 000 compounds, and after running a protocol of docking, 100-1000 candidate molecules remain to post-analysis. The latter is an essential procedure since scoring functions have several drawbacks and non-ligands might be prioritized first than true ligands, which is not desirable. Thus, the final step in SBVS strategies is a thorough manual process of hit selection, in which binding modes of hundreds of top-scoring compounds are inspected in molecular graphics programs. Nonetheless, the identification, prioritization, and automatic selection of promising HITs is still an open problem in the SBVS field. Bearing this in mind, we propose MOTIF, a novel hashed interaction fingerprint that encodes interactions on molecular complexes both as binary or count fingerprint. Differently from other hashed fingerprints that are usually black-boxes, in which one has to design its own methods to interpret what each bit represents, MOTIF already provides several features to make the analysis straightforward and out-of-the-box. Moreover, as an effort to validate and illustrate the applicability of MOTIF, we selected a recently published library consisting of 138 million molecules docked against Dopamine D4 as our case study. Herein, our goal was to train different machine learning models to reproduce Dock scores. In this scenario, we showed that MOTIF outperforms state of the art methods with an R-squared of 0.47.

Funding:

,

The relationships between variability, architecture and mutation co-occurrence in the HIV-1 integrase: implications of Raltegravir treatment.

Lucas de Almeida Machado, Ana Carolina Ramos Guimarães

Fiocruz

Abstract

The integrase of HIV-1 is one of the primary targets in antiretroviral therapy. This enzyme is responsible for integrating the viral DNA into the host genome, a crucial step in the HIV-1 replication cycle. The integrase inhibitor Raltegravir (RAL) has been widely used in antiretroviral therapy; however, the emergence of RAL-resistant HIV-1 strains has become a worldwide problem. Here, we compared the variability of each position of the HIV-1 integrase sequence in clinical isolates of RAL-treated and drug-naïve patients by calculating their Shannon entropies. We also built tridimensional models of the HIV-1 integrase and a mutation co-occurrence network. The relationship between variability, architecture, and co-occurrence was investigated. It was observed that positions bearing major resistance-related mutations are highly conserved among non-treated patients and variable among the treated ones. The integrase structure showed that the highest-entropy residues are in the vicinity of the host DNA, and their variations may impact the protein-DNA interface. The co-occurrence network and structural analysis support the hypothesis that the resistance-related E138K mutation compensates for mutated DNA-anchoring lysine residues. Our results reveal patterns by which the integrase adapts during the RAL therapy. This information can be useful to rethink the drugs currently used or to guide the development of new ones.

Funding: CAPES

,,,

A Structural bioinformatics approach for Functional characterization of Treponema pallidum subspecies hypothetical proteins

Arun Kumar Jaiswal, sandeep tiwari, Vasco A de C Azevedo, Siomar de Castro Soares

Universidade Federal do Triângulo Mineiro

Abstract

In the 21st century, there is yet an unsatisfactorily high worldwide rate of Sexually transmitted infections (STIs), around the globe, in excess of a million STIs are obtained each day. As per World Health Organization (WHO), in 2016 there were an approximated 376 million new cases of the four treatable STIs-Chlamydia, Gonorrhea, Syphilis and Trichomoniasis and among those 6 million cases has been accounted for just of syphilis and more than 11 million new infections of syphilis occur each year. The gram negative medically important spirochete bacterium called Treponema pallidum (highly virulent bacterium) subspecies pallidum is causative agents of Syphilis. There are three progressively known species from same Genus are causes human treponemal infections, for example, Treponema pallidum subspecies pertenue that causes yaws, Treponema pallidum subspecies carateum causes pinta and Treponema pallidum subspecies endemicum causes bejel or endemic (Nonvenereal-transmission of sickness without Sexually contact) syphilis. The molecular mechanisms mainly pathogenesis of Treponema pallidum are not well known. This is because Treponema pallidum can not be cultured in laboratory, naturally fragile behavior and phylogenetically no conventional virulence factor homologous with other pathogens makes it difficult to work with. Actually almost 30% of its proteins coding genes are not functionally known and characterized as hypothetical protein. In this work we applied a structural bioinformatics technique using Phyre2 web-based homology modelling (tertiary structure homology modelling) to better understand and annotate the hypothetical proteins based on proteome wide scale of Treponema pallidum subspecies level. In this work we used genome of each subspecies (complete genomes of Treponema pallidum subspecies pallidum, Treponema pallidum subspecies endemicum and Treponema pallidum subspecies pertenue). We found 297 (30%) of protein coding genes were hypothetical in Treponema pallidum subspecies pallidum strain Nichols and after comparing these hypothetical protein coding genes with Treponema pallidum subspecies endemicum and Treponema pallidum subspecies pertenue, 253 (26%) of protein coding genes were common. The 297 out of 969 Treponema pallidum subspecies pallidum strain Nichols protein modeled with Phyre2-based tertiary structure modeling with high-confidence score which were assigned as hypothetical proteins with no functions in published proteomes. Hypothetical proteins modeled in this work with high-confidence were predicted that showed remarkable structural similarity with proteins that experimentally confirmed to be required for virulence in other pathogens. Significantly, our tertiary structure modeling approach was also able to predict structural models based on functionally annotated templates for over of all hypothetical T. pallidum proteins, which will help to better understand the structure-function relationships and fundamental molecular mechanisms of T. pallidum pathogenesis at subspecies level.

,,,

Molecular modeling and pharmacophore based virtual screening of The Nicotinic acetylcholine receptor of Halyomorpha halys

Beatriz Pereira do Nascimento, Fabrício Santos Barbosa, T. S. Melo, Bruno Silva Andrade

Universidade Estadual do Sudoeste da Bahia, Brazil

Abstract

The irrational use of fertilizers as a pest control treatment has become increasingly a potential problem for the industrial agricultural sector. In addition, nicotinoid resistant pests has been increasing over the years, and for this reason searching alternative compounds for controling and erradication o these pests is crucial for crop production next years. Halyomorpha halys is popularly known as brown marmorated stink bug, and it spreads in soybean crop, damaging most of the grains in formation, as well as is responsible for the reduction in seed yield and quality. The aim of this work was to construct the nAChR 3D structure of H. Halys as well as perform a virtual screening study in order to find new compounds which can complex and inhibit this receptor. The nAChR 3D structure was modeled using homology modeling approach by SWISS MODEL software. Known nAChR inhibitors were used to perform a pharcophore alignment with Pharmagist (http://bioinfo3d.cs.tau.ac.il/PharmaGist/) and after it was submited to ZincPharmer (http://zincpharmer.csb.pitt.edu/) for seaching for pharmacophore-like ligands in ZINC database. In a second step we docked 1.000 selected molecules into the nAChR active site AutoDock Vina software. The five complexes with best enegy affinity values were submited to AMBER 14 package for Molecular Dynamics simulations.

Funding: 1 – Programa de Pós-graduação em Química. Universidade Estadual do Sudoeste da Bahia, Brazil. 2- Departamento de Ciências e Tecnologia (DCT). Universidade Estadual do Sudoeste da Bahia, Brazil 3 Laboratório de Bioinformática e Química Computacional – LBQC. Universidade Estadual do Sudoeste da Bahia, Brazil

”

INSIGHTS OF THE IRF1 DNA BINDING DOMAIN: MODELING AND DYNAMICS OF ITS 3D STRUCTURE

Cinthia Caroline Alves, Eduardo Antônio Donadi, Silvana Giuliatti

Ribeirao Preto Medical School

Abstract

Interferon regulatory factor 1 (IRF1) is a member of a closely related proteins referred as the IRF family, which encodes a transcriptional factor (TF) responsible for the transcriptional regulation of several interferon-inducible genes involved in innate and adaptive immunity. This protein contains a DNA-binding domain (DBD) in its N-terminal region which forms a helix-turn-helix motif and recognizes the interferon-stimulated response element (ISRE) in the promoters of target genes. Mutations at its DBD was associated with cancer development, such as gastric cancer. Then, in the present study, we investigate the structure and dynamics of the IRF1 DBD to better understand its role as a transcription factor. In the Protein Data Bank (PDB - https://www.rcsb.org/), DBD crystal of the IRF1 TF (PDB code: 1IF1, with 3 of resolution) presented missing atoms in side chains, which does not allow protein structure dynamics study. Then, this crystal was used as template to perform homology modelling of human IRF1 DBD which encompasses the residues 5-113 (UniProt code: P10914 - http://uniprot.org/) using MODELLER 9.14 to generate 5 models. These models were evaluated according their stereochemical properties quality (PROCHECK) and visual analysis (PyMOL2.0) comparing template tertiary structure with homology modeling models. The best quality model was used for molecular dynamic simulations of 100 ns using GROMACS 5.1 to analyze the protein behavior in solution. Homology modeling allowed satisfactory IRF1 DBD prediction models in comparison to the template: quality torsion angle assessment of the best predicted model presented 85.4% of the residues in the core region of phi-psi torsion angles, while template presents 57.6% of the residues in allowed regions, both of them presented a unique residue in disallowed regions of phi-psi torsion angles. The best quality model showed the lowest root-mean-square deviation (RMSD) of 0.325 after model-template alignment. After energy minimization and equilibrium steps of the chosen model, the molecular dynamic simulation was done. In general, the predicted model presented low conformational deviation over time with the highest fluctuations observed in loops and terminal regions. In conclusion, homology modeling generated a satisfactory 3D structure of the IRF1 DBD, which can be very useful for in silico analysis as template to model DBD and protein-DNA interaction studies.

Funding: CAPES, CNPq

,,

Molecular modeling and pharmacophore based virtual screening of Sterol 24-C-methyltransferase from Leishmania brasiliensis

Fabrício Santos Barbosa, Tarcisio Silva Melo, Bruno Silva Andrade

Universidade Estadual do Sudoeste da Bahia, Brazil

Abstract

Acording Brazilian Ministry of Health, Leishmaniasis is described as one of the most important neglected diseases of Brazil, as well as in other 12 Latin American countries. Leishmania braziliensis Vianna is responsible for causing the tegumentary form of leishmaniasis which generates cutaneous injuries by immune cells destruction during its binnary division. The enzyme 24-sterol C-methyltransferase (EC: 2.1.1.41) belongs to the transferase family and is responsible for catalyzing the transfer of methyl group in reactions for ergosterol synthesis in order to maintain membrane fluidity and permeability. The aim of this work was searching for potential inhibitors for 24-sterol C-methyltransferase addressing to block ergosterol production, as well as with minimum toxicological effects for the hosts. In a first step we used a molecular homology modeling approach using MODELLER software, for obtaining an 3D model of this enzyme. Using the AMBER14 package, we subjected the 3D model to 20.000 cycles of energy minimization followed by 10 nanosseconds of molecular dynamics. Additionally, we performed a pharmacophore based virtual screening using as start points known drugs with leishmanicidal activity – in this step we used PharmaGist (http://bioinfo3d.cs.tau.ac.il/PharmaGist/php.php) for generating the sdf output molecular aligment. Furthermore, we subjected the molecular alignment to ZincPharmer (http://zincpharmer.csb.pitt.edu/pharmer.html) for searching pharmacophore-drug-like ligands, which returned 3025 molecules. All molecules selected in pharmacophore studies were used for molecular docking calculations by AutoDock Vina software. Considering punctuation criterion as well as stereochemical and binding characteristics we selected the 30 best ligands with affinity energies below -12.0 Kcal/Mol. Molecular interactions and 2D interaction maps were generated with PyMOL 2.1.1 and Discovery Studio, respectively. In a further step we will perform molecular dynamics of 50 nanosseconds for the 10 best complexes.

Funding:

’’’’’

Novel HMG-CoA reductase inhibitors development by integrating dyslipidemic patients' genetic studies and molecular modelling

Glaucio Monteiro Ferreira, Victor Fernandes de Oliveira, Thales Kronenberger, Rosário Dominguez Crespo Hirata, Mario Hiroyuki Hirata, Fausto Feres

Laboratory of Molecular Research in Cardiology (LIMC), Dante Pazzanese Institute of Cardiology, São Paulo, Brazil.

Abstract

Dyslipidemias are a group of functional disease caused by any alteration in lipid metabolism, resulting modifications in plasma of lipoproteins. The most important lipoprotein related with high risk to develop atherosclerosis is low-density lipoproteins, that is treated mainly by statins, the first-choice pharmacological therapy, a potent inhibitor of 3-hydroxy-3-methyl-glutaryl-coenzyme A reductase (HMGCR). Previous study we identified functional variants of HMGCR, whose structural conformations modify the molecular interactions with HMG-CoA/NADPH/statins, thus contributing to the treatment failure. This data can explain in part a decreased treatment efficiency. This study aims to shed a light on the HMGCR-resistance phenomena from a structural point of view. Therewith, in order to obtain new HMGCR-inhibiting drugs, the use of strategies rational drug planning integrating genetic knowledge, pharmaceutical chemistry, pharmacology, biochemistry and molecular modelling could contribute to reduce the current pandemic scenario of cardiovascular diseases and reduce public costs. 3D structures of the proteins containing HMGCR variants generated by homology modeling method. The HMGCR crystallographic structure that used was deposited in the Protein Data Bank (PDB: 1HWK), which was used to identify important regions of the active site, disorganized regions and conserved regions of the HMGCR, the global alignment was carried out to detect these regions. In order to validate the model, the quality of the model evaluated in comparison to the crystallographic structure by Ramachandran plots. Variants were identified in HMGCR gene in our sequencing database at the Dante Pazzanese Institute of Cardiology (IDPC) and compared with the literature (IDPC: rs12916; rs59083; rs5909 and rs17238554; rs5908; rs2241402 and rs3846662). It is noteworthy that only rs5908 (Ile638Val) is exonic and missense variant. However, its clinical significance remains to be fully characterized. Ultimately, the understanding, at the molecular level, of these mutations and their functional impact on the statin interaction. For instance, the rs5908 polymorphism, which is contained in the interface between subunits HMGCR, may have further structural implications since it is opposite to the statin binding site. Patients who present this variant demonstrate clinical data of cholesterol fractions between CT: 250 mg/dL; HDL: 50 mg/dL e LDL: 210 mg/dL (sequencing database at IDCP). The results will overall contribute to the scientific knowledge in the area of drug discovery as well as in the field of pharmacogenomics, highlighting the importance of understanding the genetic mechanisms and integrating information for the discovery of new drugs, not only for the treatment of dyslipidemias but also for the treatment of other diseases.

Funding: FAPESP project n^o 2019/06172-4

,,,,,

PCSK9 three-dimensional reconstruction by homology modelling and new LDL receptor interaction regions

Vitor Galvão Lopes, Victor Fernandes de Oliveira, Thales Kronenberger, Mario Hiroyuki Hirata, Rosário Dominguez Crespo Hirata, Glaucio Monteiro Ferreira

Laboratory of Molecular Biology applied to Diagnostic (LBMAD), Department of Pharmacy, Faculty of Pharmaceutical Sciences, University of São Paulo, São Paulo, SP, Brazil.

Abstract

Dyslipidemias are a group of functional disease caused by any alteration in lipid metabolism, resulting modifications in plasma of lipoproteins. The most important lipoprotein related with high risk to develop atherosclerosis is low-density lipoproteins, that is treated mainly by statins, the first-choice pharmacological therapy, a potent inhibitor of 3-hydroxy-3-methyl-glutaryl-coenzyme A reductase (HMGCR). In this way, proprotein convertase subtilisin/kexin type 9 (PCSK9) is an enzyme that cleaves low-density lipoprotein (LDL) receptor and therefore it controls cholesterol homeostasis. Some specific mutations cause gain-in-fuction of PCSK9, which result in increased cholesterol levels in blood. In this way, the PCSK9 has been used as a target to develop cholesterol lowering terapies. This work aimed to build a tridimensional model of Proprotein convertase subtilisin/kexin type 9 (PCSK9) protein to study the interaction regions with the LDL receptor, an important molecule in cholesterol homeostasis. The PCSK9 3D structure (Proprotein convertase subtilisin/kexin type 9) were constructed by homology modelling method. In this way, was used the crystallized structure deposited in Protein Data Bank (PDB: 2P4E) to identify the catalytic site regions of interest, between disorganized or conserved regions of PCSK9. The global alignment was validated by Z-score and Ramachandran graphics. In the 3D model of PCSK9, it was possible to identify regions had not been elucidated in the crystallographic structure (Figure 1). The regions reconstructed by the homology modeling method are mainly found in the PCSK9 catalytic domain (in yellow), where interactions with the LDL receptor occur. The C-terminal domain (in green) is responsible for the structural stability of PCSK9. The pre-domain (in red) is the self-cleaved region by PCSK9 itself and is not part of the tertiary structure that interacts with the LDL receptor. In the catalytic domain, amino acids (Asp168, Glu169, Tyr170, Gln171, Pro172, Pro173 and Asp174) that are not found in the PDB:2P4E crystallographic structure can be observed, possibly due to the limitation of the crystallography method in predicting very labile regions. The Ramachandram graph presented 0.1% (Gly68) of amino acids outliers, demonstrating robustness of the constructed model. Three-dimensional reconstruction makes it possible to understand the structure of PCSK9 and to identify new regions of interaction with the LDL receptor that may be important for the development of new drugs.

Funding: FAPESP project n^{o} 2019/06172-4

,

A new bioinformatics pipeline to identify schistosomiasis vaccine candidates from a phage-display assay

João Vicente de Morais Malvezzi, Sergio Verjovski-Almeida

Instituto Butantan

Abstract

Some species of Schistosoma spp. are haemoparasites that cause schistosomiasis, affecting more than two hundred million people worldwide. In Brazil, the etiological agent is S. mansoni. Since the publication of its transcriptome in 2003, little success has been achieved in tests of vaccine candidates, and new methods for recognizing new targets are required. In the present study, we have applied a phage immunoprecipitation followed by sequencing (PhIP-Seq) and employs self-healed rhesus macaques antibodies against S. mansoni to capture antigenic peptides from a phage-display library constructed by us with synthetic oligonucleotides that encode all fragments of all S. mansoni proteins having known sequences, with the aim of identifying new epitopes for a vaccine against schistosomiasis. We developed a bioinformatics pipeline to analyze the data that will be obtained by PhIP-Seq. For each FASTQ the pipeline applies steps of quality control, preprocessing, read mapping against the oligonucleotides library, and count the reads mapped to each oligonucleotide sequence to generate a table of read counts for each sample. The count tables are processed with a script in R which will identify enriched peptides for each sample. To find enriched peptides in the immunoprecipitated phages compared with peptides abundance in the total (non-precipitated), three scripts were created using different statistical approaches:1) Zero-Inflated Generalized Poisson (ZIGP) distribution, 2) Negative Binomial (NB) distribution, and 3) Z-score. To test these scripts, we used a real PhIP-Seq dataset from a phage-display library of human-virus protein fragments containing 96, 179 peptides screened with 10 human serum samples with two replicates each. The ZIGP script identified 1, 757 peptides as enriched, whereas NB script and Z-score script identified 1, 243 and 2, 369, respectively. ZIGP and NB scripts identified 1, 126 peptides in common, while only 30 peptides were enriched by the three methods. These results show that the NB is more stringent, returning fewer peptides in comparison to the other methods. Noteworthy is that a more robust group of enriched peptides is acquired when using the intersected result between NB and ZIGP when compared with those enriched peptides returned by Z-score method. Given that, this pipeline will be applied to search for antigens of S. mansoni recognized by the antibodies from twelve self-healed rhesus macaques sera. We expect to find groups of macaques that developed antibodies for different antigens or even the same antibodies at different times after infection and healing. Probably those proteins targeted by the macaques' antibodies are critical for the survival of the parasite. Thus, we expect to identify several new parasite antigens candidates that can be tested in future studies to generate a vaccine against schistosomiasis.

Funding: FASPESP

””

Analysis of Affinity and Selectivity of Novel Inhibitors of Polyketide Synthase 13 of Mycobacterium tuberculosis by Molecular Dynamics Simulation and Binding Free Energy Calculations

Jorddy Neves Cruz, João Marcos Pereira Galúcio, Paulo Henrique Taube, Kauê Santana da Costa, Eloisa Helena de Aguiar Andrade

Museu Paraense Emilio Goeldi, Botany Coordination. Adolpho Ducke Laboratory, Belém, Pará Brazil.

Abstract

Polyketide Synthase 13 (Pks13) is an essential enzyme that forms mycolic acids, which are critical for viability and virulence of Mycobacterium tuberculosis. Pks13 performs the final assembly step of the mycolic acid synthesis, i.e., the Claisen-type condensation of a C26 a-alkyl branch and C40–60 meromycolate precursors. In the present study, we investigated the binding mode, the affinity, and selectivity of novel inhibitors, Tam5 and Tam6, against the Pks13 binding pocket by molecular dynamics simulation (MD) and binding free energy calculations. Our analyses showed that all Pks13-inhibitors systems reach the stabilization after 30 ns of MD, exhibiting for protein backbone an average RMSD value of 1.61 and 1.59 and the inhibitors also showed a high affinity to the residues of binding pocket exhibiting the following energies (?Gbind) -46.26 $\pm$ 0.07 kcal.mol-1 and -36.52 $\pm$ 0.05 kcal.mol-1, respectively. Ligand pairwise per-residue energy decomposition analysis showed that Ser1636, Tyr1637, Asn1640, Ala1667, Phe1670, and Tyr1674, exhibited the most energetic contribution for ligands stabilization in Pks13 binding pocket. These preliminary results will be useful to further in silico studies that aim to develop novel analog inhibitors with improved selectivity and affinity against Pks13 binding pocket.

Funding:

,,,

Analysis of Structural Evolution of FT and TFL1 Proteins of Angiosperms

Deivid Almeida de Jesus, Darlisson Mesquita Batista, Kauê Santana da Costa, Thiago André

UFOPA

Abstract

Flowering Locus T/Terminal flowering T1-like (FT/TFL1) genes stimulating and suppressing flowering in angiosperms. In the present study, the ancestral sequences were reconstructed by phylogenetic analyzes and then their structures and evolutionary relations were analyzed. First, the coding sequences were obtained from GenBank using the BLASTn tool and the Arabidopsis thaliana sequence was used as a reference. The nucleotides sequences were aligned in the MUSCLE and the most verisimilar evolutionary model (GTR+I+G) was retrieved in jModelTest 2 program and validated with 1, 000 bootstrap repetitions. For phylogenetic inferences, we used a Bayesian approach in BEAST program, using Pinidae subclass species as an outgroup. The ancestral sequences of angiosperms, monocotyledons, tricolpades, asterids, brassicales (FT/TFL1) and non-brassicales (TFL1) were obtained by maximum likelihood method available in MEGA7. The FT/TFL1 structures were obtained by comparative modeling in Modeller program using as a template the crystallographic structure of the FT protein of A. thaliana (1WKP/1WKO, chain A). Then, molecular dynamics (MD) simulations were performed using the Amber16 package to analyze possible conformational changes in the modeled structures. The stereochemical quality of each model was optimized by ModRefiner program. Finally, the models were validated using the Ramachandran plot, which evaluates the stereochemical quality; and by the QMEAN plot, which evaluate the energy profile. In addition we performed a mutational analysis of the protein structures using FoldX. The modeled proteins showed an RMSD-Ca = 1.6 and at least 88% of the residues in favorable regions of the Ramachandran plot. The phylogenetic relations of proteins demonstrated that FT/TFL1 proteins do not show the same evolutionary divergence found in angiosperms, and speciation and selection acted differently in these floral proteins. There was little structural variation in FT/TFL proteins throughout the evolutionary history of angiosperms, as shown by the RMSD values obtained after MD simulations. Induced mutations have shown that proteins undergo high destabilization in various residues, as well as, in the region of the fourth exon, which is important for protein activity. Considering the importance of the floral activation, we concluded that the evolutionary conservation of the FT/TFL1 structures must evolve under stabilizing natural selection.

Funding:

,,,,,

A DATABASE OF GLUCOSE-TOLERANT β-GLUCOSIDASES

Leandro Liborio da Silva Matos, Diego César Batista Mariano, Naiara Pantuza, Luana Luiza Bastos, Letícia Xavier Silva Cantão, Raquel Melo Minardi

Universidade Federal de Minas Gerais - UFMG

Abstract

β-glucosidase (BGL) is an important enzyme for the production process of lignocellulosic biofuels. It is responsible for cellulose degradation in the glucose used in the fermentation process. However, the literature has described that most BGLs are inhibited by glucose. BGL has been classified into several subfamilies of glycosyl hydrolases, such as GH1, GH3, GH5, GH9, GH30, and GH116. Studies have shown members of family GH1 are more efficient for cellulose degradation in biofuel due to their higher resistance to product inhibition. Recently, a database of glucose-tolerant β-glucosidases, called Glutantbase, has been proposed for understanding the impact of point mutations in distinct structures. Glutantbase contains data of multiple alignments of GH1 enzymes obtained from Clustal Omega, catalytic residues, secondary structure, and extrapolated mutations for more than 3, 500 GH1 β-glucosidases sequences extracted from UniProt that were modeled by reference using MODELLER. However, many aspects of some mutations have not been established. For instance, the mutations E96K and M416I in Bacillus polymyxa β-glucosidase, that have been reported to thermostability. A computational way to try to understand better the impact of these mutations in BGL structure is through coevolution networks. In this study, we proposed to use coevolution networks to try to understand better the impact of mutations. To determine the coevolution network, we can use the PFSTATS tool. In addition, graph modeling can be used to try to understand the importance of correlated residues with many other residues. Finally, alanine scanning can be used to obtain the importance of the residues highly statistical coupling. We hope that the results obtained in this work may shed light on a new generation of second-generation biofuels. Glutantbase is available at <http://bioinfo.dcc.ufmg.br/glutantbase/>.

Funding:

,,,

Computational Identification of Orthologous Proteoforms between Human and Murine

Letícia Graziela Costa Santos de Mattos, Esdras Matheus Gomes da Silva, Vinicius da Silva Coutinho Parreira, Fabio Passetti

Laboratory of Gene Expression Regulation, Carlos Chagas Institute, Fundação Oswaldo Cruz (Fiocruz), Curitiba, PR, Brazil

Abstract

The advent of next-generation sequencers in Transcriptomics and mass spectrometry in Proteomics has resulted in a large volume of available data. These datasets became integrated into several Bioinformatics studies in an area called Proteogenomics. Thus, the fraction of messenger RNAs (mRNA) that are effectively translated into proteins has been deeply studied. Alternative splicing (AS) is a molecular event that may occur during mRNA maturation in more than 95% of human genes. AS might produce several mRNA isoforms that can change amino acids sequence, and consequently different proteoforms. In this context, the main goal of this project is to incorporate algorithms to our methodologies that allow us to identify orthologous proteoforms between humans and murine. For this purpose, we used transcriptome data from the Ensembl project to create a sequence repository using the ternary matrices methodology, which was developed by our research group. This customized sequence repository, created for human and murine datasets, was used to identify AS isoforms. In the human datasets, we were able to identify 22, 242 splicing variants in 61, 122 genes. According to the Ensembl Transcript Support Level (TSL) parameter, 41, 080 were ranked as reliable and 181, 382 with lower reliability. The gene that presented more AS variants was MAPK10, with 192 transcripts. Other human genes with known splice variants such as BCL2L1, KLF6, and TMP2 were also detected in our database. In the mouse transcriptome datasets, we found 126, 679 splicing variants in 43, 976 genes. From these variants, 43, 985 were classified as reliable and 82, 694 with lower reliability, based on TSL. Henceforth, for our next steps, we aim to identify expressed orthologous proteoforms using RNA-Seq data and proteomic shotgun data from human and murine healthy tissues. Additionally, we intend to select a list of proteoforms for experimental validation.

Funding: Coordenação de Aperfeiçoamento de Pessoal de Nível Superior; Conselho Nacional de Desenvolvimento Científico e Tecnológico

””

Taxonomic classifier for β-glucosidase enzymes based on structural signatures

Letícia Xavier Silva Cantão, Luana Luiza Bastos, Leandro Liborio da Silva Matos, Marcos Augusto dos Santos, Raquel Melo Minardi

Universidade Federal de Minas Gerais - UFMG

Abstract

β-glucosidases are important enzymes for catalysis of hydrolysis of β-glucosidic bonds and can be applied in the field of biotechnology. They are involved in biofuel generation processes through the production of fermentable sugars. These enzymes can be found in many types of organisms such as bacteria, archaea, and eukaryotes. Structural signatures can be defined as patterns of distance between protein atoms. An example of a technique for calculating these distances is aCSM all software, where the calculation is performed for all atoms against all. The use of fresh new linear algebra techniques to construct logistic regression models estimates the probability associated with the occurrence of a given event in the face of a set of explanatory variables. The goal of this work is to construct a taxonomic classification model based on the structural signatures of β-glucosidases enzymes, considering the domain taxonomic level to which they belong, through the use of linear algebra and modified logistic regression. For the construction of the model, we used 162 PDB files of β-glucosidase enzymes from bacteria, archaea, and eukaryotes. From these PDBs files, we generated structural signatures using aCSM software, which resulted in a distance matrix with 7200 attributes and 162 entities. From this matrix, we build three datasets, one for each domain. To construct the classifier we divided each dataset into 80% for training and 20% for testing, applying linear algebra and modified logistic regression to calculate the probability of each entity belonging to its respective domain. We evaluated the constructed model through the (fa) - harmonic mean and the area under the ROC curve. The classifiers generated by the modified logistic regression models presented good results, being that the average harmonic average and the average area on the ROC curve for the archaea, bacteria, and eukaryotes enzyme classifiers were respectively: 0.88 / 0.89, 0.90 / 0.90, 0.89 / 0.88. The classifier proved to be efficient and promising for determining the taxonomic domain of β- glucosidase enzymes based on their structural signatures.

Funding:

,,,

Interaction Between TNF and SVMPs of PI Class: Molecular Modeling and Docking at a Glance

Luana Luiza Bastos, Letícia Xavier Silva Cantão, Leandro Liborio da Silva Matos, Raquel Melo Minardi

Universidade Federal de Minas Gerais - UFMG

Abstract

TNF (Tumor Necrosis Factor) produced mainly by monocytes, is a proinflammatory cytokine that plays an important role in modulating inflammatory responses and host defense mechanisms. At increased levels such cytosine is closely related to degenerative diseases such as rheumatoid arthritis. Metalloproteases are enzymes characterized by their zinc catalytic structure at their active site, snake venom metalloproteases (SVMPs) are responsible for inducing haemorrhage and disturbances in the prey blood coagulation cascade. However, certain SVMPs do not have hemorrhagic activity (PI class SVMPs), thus having effects such as inhibition of platelet aggregation, apoptosis induction and pro anti-inflammatory activities. Previous studies carried out in silico and in vitro with the enzyme BmooMP-alpha-I isolated from snake venom Bothrops moojeni a metalloprotease class PI have demonstrated its ability to directly inhibit TNF in immunopathologies like colitis.The aim of this study is to perform in silico studies to elucidate the interaction between tumor necrosis factor and metalloproteases class PI present in snake venom. For this will be carried out protein-protein interaction dockings of TNF and SVMPs, structural alignments and molecular modeling will be performed. As result, it is expected to be obtained Ligand Root Mean Square Deviation LRMS = 5.0 or Interactive Root Mean Square Deviation IRMS = 2.0 . At the end of this project we hope to contribute to the understanding of the interaction between TNF and SVMPS class PI and their therapeutic applications.

Funding:

,,

IDENTIFICATION AND MEASUREMENT IN SILICO OF PROTEIN TUNNELS AND LIGATION POCKETS OF THE CRY3BB1 INSECTILE TOXIN.

Luis Angel Chicoma Rojas, Renato Farinacio, Eliana Gertrudes de Macedo Lemos

Private University Antenor Orrego & Paulista State University, Jaboticabal Campus

Abstract

The rising increase in the discovery of toxins with potential for pest control presents challenges in the understanding of their physical-chemical characteristics and structural attributes, which makes it difficult to select the best candidates for this function. However, thanks to the development of bioinformatics tools it is possible to analyze toxic molecules such as Cry proteins, produced by the Bacillus thuringiensis bacteria, efficiently and with low costs, therefore, the objective of this work is to identify and size the different tunnels and protein pockets with ligation powers of the Cry3Bb1 toxin. The NCBI database was used to search for the sequence of Cry3Bb1 (GenBank: AAA22334.1). In the three-dimensional modeling, the SWISS-MODEL online server was used, and the Pymol2.0 program for the visualization and manipulation of the structure. Likewise, the Electrostatic Surface of the toxin was calculated using the APBS program (Adaptative Poisson-Boltzman Solver). The detection and dimensionality of the Cry3Bb1 protein pockets was performed with the D3Pockets server and the CASTp program; finally, to locate and characterize (hydrophobicity and polarity) the protein tunnels was used the MOLE 2.5 program. The results show that the Cry3Bb1 protein has 14 pockets, where the major and minor have an area and volume of 501, 817 and 5, 998 (SA); and 312, 233 and 0.542 (SA), respectively. In addition, 7 tunnels were identified in the structure of the cry3Bb1 protein, where the largest has a length of 20.27 and the smallest of 5.25. Taking into account the results obtained from the structural analysis of the toxin, the use of bioinformatics tools demonstrates great potential in understanding the architecture and properties of insecticidal interest's molecules.

Funding:

”

COMPARATIVE ANALYSIS OF THE THREE-DIMENSIONAL STRUCTURAL OF THE CRY23AA1 AND CRY51AA1 INSECTILE PROTEINS USING BIOINFORMATIC TOOLS.

Luis Angel Chicoma Rojas, Renato Farinacio, Eliana Gertrudes de Macedo Lemos

Private University Antenor Orrego & Paulista State University, Jaboticabal Campus

Abstract

Bacillus thuringiensis has been widely used as a bioinsecticide or in the genetic transformation of plants because of its ability to produce a wide variety of active toxins, such as Cry proteins. The use of computational methods for the study of the three-dimensional structures of Cry proteins allows a better understanding of their functionality and classification, for this reason, the objective of this work is to make a comparison between the tertiary structures of the Cry23Aa1 and Cry51Aa1 proteins, and their respective amino acid sequences, to identify the similarity between said biomolecules.The NCBI database was used to obtain the amino acid sequence of the Cry23Aa1 (GenBank: AAF76375.1) and Cry51Aa1 (GenBank: ABI14444.1) proteins. The SuperPose 1.0 program was used to perform a sequence alignment between proteins and identify their degree of identity and similarity. To obtain and display / edit the three-dimensional structures of Cry23A1 and Cry51Aa1, the SWISS-MODEL server and the Pymol 2.0 program were used, respectively. The electrostatic surface of both proteins was calculated with the APBS program (Adaptative Poisson-Boltzman Solver). Finally, to generate a matrix graph, a structural overlap was made between Cry23Aa1 and Cry51A1 with the Distance Difference Matrix (DDM) of SupePose 1.0 and structural alignment commands of Pymol 2.0. The results showed that at the amino acid sequence level the Cry23Aa1 and Cry51Aa1 proteins have a score, identity and similarity of 127.5, 20.8% (67/322) and 34.2% (110/322), respectively. At the level of three-dimensional structure, a Root-Mean-Square Deviation (RMSD) of 1.85 was identified, indicating a high similarity. Through computational biology tools it is possible to understand the differences that exist at the structural level of different molecules that are of interest, allowing a better classification of these.

Funding:

,

PFMutStats: A new method for describing missense mutations by Annotations, Conservation, Coevolution, Interactions and Structural Feature

Marcelo Querino Lima Afonso, Lucas Bleicher

Universidade Federal de Minas Gerais

Abstract

Predicting the functional effect of missense mutations in proteins is essential for understanding the mechanisms of several diseases, and for the rational design of proteins for various applications. In this work we aimed to develop a highly integrative methodology for analysing the possible effects of a given mutation in the functions exerted by a protein. This work is based on analysing Multiple Sequence Alignments of Protein Domains from the Pfam database. Several mutation descriptors are already implemented in a Web Application such as: amino acid frequency before and after mutation, position conservation, UniProt Protein Families annotations, residue interactions established before and after a mutation, centrality metrics related to the residue interaction network of a protein, residue depth and secondary structure assignment, and Ramachandran Distribution analysis before and after the mutations. Various dynamic plots were developed in the D3.js library in order to illustrate and interpret the descriptive results. For the Ramachandran Distribution analysis a Database of reference structures was assembled. All Pfam Protein Family alignments from the UniProt database were downloaded and parsed for calculating and storing the amino acid frequency and column conservation by the Jensen-Shannon entropy. Residue depth was calculated using the latest theoretical model for maximum accessibility as a reference and assignment to the surface or protein interior was based on calculating two vectors of the amino acids frequencies at either site using different assignment thresholds and looking for maximum divergence between these vectors. Residue Interactions Networks were generated by the RING Software and a given chain rank in graph centrality metrics (Betweenness, Closeness and Degree) were calculated and weighted based on the interactions energies. Other features have been planned and should be implemented in the following year.

Funding: Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES).

”

The interaction of NS5 protein with the human importin and exportin proteins

Marcos Freitas Parra, Ana Ligia Scott, Antonio Sergio Kimus Braz

UFABC

Abstract

The Denv NS5 protein is well conserved among the 4 Denv serotypes, reflecting its vital role in the replication cycle of viral RNA. In order to replicate efficiently, viruses need to block host resistance mechanisms - which is achieved through viral proteins that control host machinery leading to viral replication and blocking eventual resistance mechanism, the NS5 protein blocks the cellular response for interacts with interferon proteins. Beyond of interferon proteins, the NS5 interacts directly with a large amount of proteins, including the proteins that import and export other to the cell nucleus. The mechanisms of NS5 interaction with that proteins is not know yet. To investigate this interactions, we use a methodology that uses the normal modes and docking approach. The approach uses Normal Modes to analysis is the approximation of the system's free energy dynamics to the Hooke Harmonic potential, which presents movements of the various protein regions around an energy minimum. Using VMOD protocol, were generated structures and using the lowest energy model can be investigated through the using of the concept of lower energy in an interaction between binding proteins and the use of the extended conformational selection model - thus representing the most likely conformational model for the docking between the two proteins. After that, large-scale tests of protein-protein interaction are performed. For these tests, proteins are treated as rigid bodies by making rotational and translational motions only, we provided conformational variations of each ligand and receptor obtained by normal modes analysis, as we selected the ligand hotspots through rigid docking, using the probability of the canonical ensemble. Stability tests and the adjustment of the interfaces were calculated by molecular dynamics. Through protein-protein docking experiment, we can find a possible hot-spot for docking between KPNB1-NS5 anb XPO1-NS5. This binding site is being confirmed through molecular dynamics, until stabilization of the system.

Funding: UFABC, CAPES

”

Can we predict protein essentiality based on their pyshico-chemical features ?

Mauricio Lopes Casagrande, Ney Lemke, Marcio Luis Acencio

Abstract

The way genes organize and behave during the life of a living being hints that there may be a set of genes essential to life and reproduction, acknowledged as essential genes. These genes show high evolutionary conservation rate when compared to genomic media and stay relatively preserved through time, making it possible to find homology even on distant species. The process of discovering essential genes in vitro requires gene expression modulation techniques such as single-gene knockout, RNA interference, conditional knockout or CRISPR, and has always been an extensive, laborious effort. Nowadays it is possible to shorten this effort by computational, ortogenetic or phylogenetic approaches, but any method able to shorten this effort even more would be welcome. This work intends to present an in silico method to aid prediction of gene essentiality based only on physico-chemical features of the proteins synthesised by a gene. We applied machine learning techniques to verify if there really is some kind of relation between these features and gene essentiality and to evaluate the predicting capability of two algorithms using two distinct organisms. The knowledge about the set of essential genes of a certain organism could help improving our comprehension about the mechanisms related to the genetic base needed to sustain life, signaling new effective antimicrobial drug targets, increasing our knowledge about synthetic life or even guiding genetic therapy.

Funding:

,,,,,

Computational Approach of NSCLC markers applied to drug design against Pd-l1 and Homology Modeling by Tusc2 (FUS1)

Patrícia da Silva Antunes, Nelson José Freitas da Silveira, Levy Bueno Alves, Thiago Castilho Elias, Márcia Paranho Veloso, William Mesquita da Costa

Universidade Federal de Alfenas

Abstract

The treatment of cancer is part of public health policy in Brazil. The needs for the development of new drugs again different mechanisms of tumor cell survival, makes computational simulation essential to predict the interaction of molecules with selected targets. Non-small cell lung cancer (NSCLC) is one of the most aggressive types of cancer, with low survival rate and high metastatic capacity. In this work, we selected two NSCLC markers, Pd-l1 and Tusc2. Pd-l1 is a Pd-1 binding protein that it is normally expressed in cells of the immune system and highly expressed in tumor cells as an escape mechanism for cell death. On the other hand, Tusc2 is a tumor suppressor candidate, and its low NSCLC expression is related to the regulation of Pd-l1. Currently, drugs that block the Pd-l / Pd-l1 checkpoint are antibodies that rely on protein engineering and cause immunogenicity problems. Therefore, in the first study, we used the strategy to develop small anti-Pd-l1 molecules that block the Pd-1 / Pd-l1 interaction. In the second study, as there isnt a reported structure of Tusc2, this work focused in the homology modeling of its 3D structure that will help in further studies with this marker. In the first study, we also simulate the interaction of active compounds from the herbal medicine Euphorbia tirucalli linneau, popularly known as Aveloz and used against cancer in the popular medicine. To perform molecular docking, we used the Induced Fit Docking protocol (IFD) as implemented in the Schrodinger suite. The pharmacokinetic properties of the molecules with the best-performing docking score (GScore) with Pd-l1 were determined using the QikProt included in the suit. The best candidate to inhibit the Pd-1 shows a docking score of -12.461 kcal/mol which represent a good result for small inhibitory molecules. Therefore, this molecule is expected to be a potential drug against Pd-1. In the second study, we used the MODELLER software to build 3D models and the Refine Loops protocol for loop optimization of Tusc2 structure. We obtained a 3D model of Tusc2 with 94.3% stereochemical quality. These results help in directing the in vitro and in vivo experiments needed to construct new drugs against NSCLC.

Funding: Federal University of Alfenas

”

Optimization of SmTGR inhibitors using a Fragment-Based Drug Design (FBDD) approach.

Rocío Lucía Beatriz Riveros Maidana, Floriano Paes Silva Junior, Ana Carolina Ramos Guimarães

Laboratório de Genômica Funcional e Bioinformática, Instituto Oswaldo Cruz, Fundação Oswaldo Cruz, Rio de Janeiro, Brazil

Abstract

Schistosomiasis is a neglected tropical parasitic disease caused by trematodes of the genus Schistosoma and specifically S. mansoni in Brazil. Schistosomiasis is the most important human helminth infection in terms of morbidity and mortality. The disease occurs in areas with poor sanitation and approximately 1.6 millions individuals were infected with S. mansoni in Brazil. Praziquantel is the unique drug employed for the treatment of the disease. Although the success of the treatment, the concern about resistance is growing and the development of new drugs is urgent. Thioredoxin glutathione reductase of Schistosoma mansoni (SmTGR) is a validated drug target that plays a crucial role in the redox homeostasis of the parasite. A crystallographic screening was held to detect the ligation of small fragments to SmTGR. 32 fragments were found to be ligated around 8 sites of the protein. After the search of analogues molecules and the study of the site of activity of the fragments, six fragments was found presenting an inhibitory activity against SmTGR acting in an allosteric site of the protein, located in the NADPH site. In this work an in silico fragment-based drug design (FBDD) will be carry out to optimize the compounds with highest inhibitory activity in order to propose new drug candidates against S. mansoni. Cavity analisis will be held to study the binding properties of the binding site. Linking and growing approach will be apply to the optimization of the hits. Evaluation of synthetic accessibility and ADMET properties prediction will be performed. Finally, the selected compounds will be object of docking and molecular dynamic studies.

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

,,,

Cradle-loop barrel in Leptospira and novel GAF fusion proteins

Rodolfo Alvarenga Ribeiro, Daniela Valdivieso, Cristiane Rodrigues Guzzo Carvalho, Robson Francisco de Souza

Institute of Biomedical Sciences, USP

Abstract

Leptospirosis is an infectious disease of high incidence in tropical regions, caused by bacteria of the genus Leptospira. The second bacterial messenger, c-di-GMP, acts on different signaling pathways that result in the regulation of virulence, mobility and biofilm formation that may be related to the infectious process. The protein encoded by the LIC_11920 gene shows DUF1577 and PilZ domains(YcgR-like and PilZ), and is a recognised as a member of the cradle-loop barrel fold, which comprehends a set of protein families that act as sensor and/or flagella structure as well as type 6 secretion system proteins, such as PilZ and YcgR. PilZ is an intracellular c-di-GMP sensor whose performance has already been related to the regulation of resistance or pathogenicity in organisms such as Borrelia, but little is known about the involvement of PilZ homologues, including LIC_11920 , on the c-di-GMP-mediated signaling pathways in Leptospira interrogans serovar Copenhageni. The DUF1577 has an unusual GAF domain in fusion with a YcgR and a PilZ domains, which could be a recent autapomorphy in the leptospiral clade. Such fusions have pointed to have relationship with diversity-generating retroelements, which could have an important role in Leptospira evolution. We intend to characterize the c-di-GMP-mediated signaling pathways in L. interrogans from the structural and functional analysis of the LIC_11920 protein, as well as to clarify the classification of the fold, in order to place the DUF1577. Along with our in-silico strategies we intend to evaluate the structure of LIC_11920 in the search for a target of pharmacological intervention in the treatment of leptospirosis.

Funding: CAPES 51/2013 CAPES 88887.357176/2019-00

,

Diversity study of small open reading frames (sORFs) of healthy and in Alzheimer's Disease brain

Saloe Bispo, Fabio Passetti

Laboratory of Functional Genomics and Bioinformatics, Oswaldo Cruz Institute (IOC), Oswaldo Cruz Foundation (Fiocruz), Rio de Janeiro, RJ, Brazil. Laboratory of Gene Expression Regulation, Carlos Chagas Institute (ICC), Oswaldo Cruz Foundation (Fiocruz), Curitiba, PR, Brazil.

Abstract

The aging of the world population is associated with the increased frequency of people diagnosed with dementias. These are responsible for the greatest burden of neurodegenerative diseases, with Alzheimer's representing approximately 60-70% of dementia cases. Computational approaches that integrate genome, transcriptome and proteome data have been developed by our research group to study human transcripts and their polypeptide products in an area known as proteogenomics. The discovery of small open reading frames (sORFs) in gene and protein databases called small ORF encoded polypeptides (sORF-encoded polypeptides) or SEPs has revealed a fundamental shortcoming in our knowledge of protein-coding genes. Some of these new sORFs have crucial biological roles in cells and organisms, which motivates the search for new sORFs. In this study, we are developing a proteogenomics approach for the identification of proteoforms in human and mouse focusing on SEPs. Thus, predictions of SEPs will be incorporated into the human and murine proteoform repository SpliceProt maintained by our group. Currently, we are using public shotgun proteomics data from healthy and AD affected brain samples, searching for new SEPs expressed under such conditions. Preliminary data have shown the presence of previously undetected microproteins in proteomics data from AD, derived from sORFs (lncRNAs and antisense).

Funding: CNPq and ICC

”

A molecular docking and ADMET study of a promising compound of the Brazilian semi arid with inhibitory potencial of IKK-B

Wagner Rodrigues de Assis Soares, Tarcisio Silva Melo, Bruno Silva Andrade

Universidade Estadual dom Sudoeste da Bahia

Abstract

The enzyme IKK-β modulates nuclear transcription factor (NF-?B) action directly affecting the transcription response of genes encoding proteins important for immune and inflammatory response. The Brazilian semi-arid compounds can be configured as scafolds for new anti-inflammatory agents. Several binders isolated from Brazilian semi arid plants were evaluated, selecting the best complexes with the IKK-β structure and their ADMET characteristics. Binder structures were designed and deposited in the Semi arid Molecule Database (SAM Database), hosted in the Laboratory of Bioinformatics and Computational Chemistry (LBQC-UESB). The enzyme structure was downloaded from PDB database (4KIK) and computational chemistry tools (Marvin Sketch, Autodock Vina, Pymol 1.7, Discovery Studio 4.0 and Osiris Property Explorer) were used to prepare the structures of the compounds for the molecular docking assay and evaluation of their ADMET characteristics. The SAM0850 binder (-6.5kcal/mol) had lower affinity energy with IKK-β, when compared to ATP (-7.3 kcal/mol), estaurosporine (-7, 7kcal/mol), GSK-7 (-9, 5kcal/mol) and higher affinity energy than acetylsalicylic acid (-5.8kcal/mol) and mesalasin (-5.5kcal/mol)/mol). The compound SAM0850 did not demonstrate toxicity in the silicon prediction, the molecular dynamics of the complexes and in vitro tests assays wiil be the next steps.

Funding:

,,,

MOLECULAR MODELING METHODS APPLIED TO THE STUDY OF Staphylococcus aureus TARGET PROTEINS

William Mesquita da Costa, Levy Bueno Alves, Nelson José Freitas da Silveira, Patrícia da Silva Antunes

Universidade Federal de Alfenas

Abstract

The application of computational methods in the rational planning of new drugs has made possible the construction of biological and chemical models that, when subjected to specific software, allow visualizing, simulating and interpreting systems involved in protein-ligand interaction. From this point of view, this work was carried out with the aim of using molecular modeling on target proteins of the Staphylococcus aureus, in order to identify molecular patterns with pharmacodynamic characteristics and inhibitory action spectrum in three enzymes present in the pathogen peptidoglycan biosynthesis, the following being: murD, ddI e uppP. From the primary sequence of murD and uppP proteins, 3D models were constructed with the help of the Modeller program. The murD and uppP sequences were obtained from the UNIPROT database, with codes P0A090 and P67391, respectively. The sequences were submitted to the PSI-BLAST algorithm in the PDB database for templates identification. Among the returned templates, the crystallographic structures 3LK7 and 6CB2, similar to murD and uppP proteins, were selected, respectively. The ddI protein was recovered from the PDB database under code 2I87 and the models were optimized. After model validation, both were submitted to molecular docking using the Glide dock-XP program, considering a virtual screening containing 1046 ligands for the identification of leading compounds. The best models generated by Modeller were murD_0332 and uppP_0703, both presenting 93.8% of amino acids residues in favorable regions on the Ramachandran plot. After optimization, the models showed good stereochemical results. The molecular docking study revealed that aminoglycoside compounds can anchor at murD, uppP and ddI binding sites simultaneously. From the structural information obtained in the virtual screening it was possible to design new molecular patterns considering ADMET pharmacokinetic aspects. Through the proposed molecular patterns it was possible to generate bioisosteres and, through new molecular docking analysis, to propose a new oral administration in silico drug.

Funding:

7 — RNA and Transcriptomics

”

Automatic identification of lncRNA transcripts using Artificial Neural Networks

Ana Beatriz Oliveira Villela Silva, Mariana Carmin, Eduardo Jaques Spinosa

Universidade Federal do Paraná

Abstract

A vast number of studies show that only 1.5% of the human genome effectively encodes proteins. Among the remaining 98.5%, a large portion is transcribed as ncRNA (non-coding RNAs), intermediate molecules that participate in the most diverse biological processes. LncRNAs (long non-coding RNAs) are a particularly recently discovered class of ncRNA which contains at least 200 nucleotides in length. Since they can be as long as (or even longer) than mRNAs (messenger RNAs) and also can contain an ORF (Open Reading Frame), differentiating between those two very functionally different types of sequences can be a challenging task. This work proposes a method to discriminate DNA sequences between mRNA and lncRNA transcripts in Human and Mouse genomes using Artificial Neural Networks. In order to train the network, 10000 (human) and 5000 (mouse) sequences from both mRNAs and lncRNAs were manually curated from the GENCODE database. The validation set for both organisms were divided into Test-A and Test-B, with Test-B's difference being the removal of lncRNA sequences that were too similar from mRNAs. A multilayer perceptron was the model adopted to classify the data. In total, nine features represented each sample in the dataset: the number and size of exons, the size of the transcript, the beginning and end of any existing ORFs, the size of the ORF found, a score calculated to represent the quality of the ORF, and the standard deviation based of the number of possible ORFs considering all possible reading frames. A grid search was performed in order to optimize the following hyperparameters: initial learning rate, number and amount of hidden layers, and the maximum number of epochs. The accuracy on both validation datasets was between 82, 36% and 93, 43% for humans and 88, 59% and 94, 18% for mice, with a precision ranging from 91, 77% and 92, 15% for humans and 93, 06% and 94, 13% for mice respectively. The recall achieved varied between 82, 36% and 83, 62% for humans and 88, 59% and 88, 89% for mice. In conclusion, a multilayer perceptron showed to be a valid classification model to discriminate lncRNAs and mRNAs, especially in Test-B, but the overall recall metric of this study can be further improved. Future work can be done using Deep Neural Network approaches in order to achieve even better results.

Funding:

,,,,,

De novo assembly and transcriptome analysis of Helicoverpa armigera feeding on natural conditions

André Ricardo Oliveira Conson, Natalia Faraj Murad, Karina Lucas da Silva Brandão, Fernando Luis Cônsoli, Celso Omoto, Marcelo Mendes Brandão

Unicamp

Abstract

The cultivation of different annual crops may provide ideal conditions for feeding and survival of lepidopteran pests presenting generalist feeding habits. With recent occurrence in Brazil, one of the most important species from Noctuidae family is Helicoverpa armigera. The main enzymes responsible for the digestive process in insects are peptidases involved in the initial digestion of plant proteins. Although plant peptidase inhibitors are an important defense mechanism against herbivory, a high tolerance is observed in H. armigera. Thus, to develop efficient ways to control pests, it is mandatory first to know which genes are involved in the digestive process and their interactions with host plants. Helicoverpa spp individuals feeding on natural conditions were collected in order to characterize differentially expressed genes associated with soybean, corn, cotton and bean diet. Total RNA from midgut was extracted and cDNA libraries sequenced (paired-end) using an Illumina Hiseq 2500. A de novo assembly of the short reads using both Mira and Trinity resulted in 240, 972 transcripts (687 bp N50) and a length of 133.35 Mb. A total of 55, 666 transcripts aligned with the SwissProt and Pfam databases. We identified 8, 429 genes differentially expressed between dietary conditions. The largest number of differentially expressed genes was obtained in the soybean versus corn feed comparison, where 1, 384 and 1, 643 genes were found down-regulated and up-regulated in soybean relative to maize respectively. Functional analysis showed that these genes are involved in biological processes like proteolysis, electron transport chain, lipid catabolic process, mRNA transport and translation. We also visualized expression patterns in important gene families, including serine protease. This is the first study of H. armigera transcriptome feeding under natural conditions and assembled transcripts are a powerful resource for future research promoting an improved understanding of the gene regulation of digestive peptidases.

Funding: Fundação de Amparo a Pesquisa do Estado de São Paulo (FAPESP) - Project 2018/19461-1, CAPES and CNPq

,,,,,

Identification of alternative splice variants in the transcriptome of Squamous Cell Carcinoma of the Cervix and Adenocarcinoma of the Cervix

Aruana F F Hansel Frose, Natasha Jorge, Patricia Savio de Araujo-Souza, Luisa Lina Villa, Laura Sichero, Fabio Passetti

Laboratory of Gene Expression Regulation, Carlos Chagas Institute (ICC), Oswaldo Cruz Foundation (Fiocruz), Curitiba, PR, Brazil.

Abstract

The uterine cervix is the initial portion of the uterus that communicates with the vaginal canal. In this region, cervical cancer can develop, which is a type of tumour that most affect women nowadays. There are many subtypes of cervical cancer: about 70% of the new cases are identified as cervical squamous cell carcinoma (SCC), 20% of the cases are adenocarcinoma (ADC), and the rest is composed by mixed tumours. Even though there are definite subtypes, the treatment administered for both SCC and ADC is the same, which compromises response and remission of ADC, the worst prognostic subtype. Since they are molecularly heterogeneous, it is still an ongoing challenge to distinguish ADC and SCC through their expression profile. The clarification of their expression profiles could complement their immunohistochemical diagnostic, help differentiate the subtypes, and possibly find new pharmacological targets differentially expressed in ADC. Since other studies have primarily focused on canonical transcripts, we propose that annotated and/or new alternative splice variants could help clarify and build the expression profile of these cervical cancer subtypes. Accordingly, our objective is to find differentially expressed alternative splice transcripts in ADC and SCC transcriptome. We obtained the RNA-Seq profile of 8 SCC and 3 ADC. After trimming using Trimgalore, mapping onto the human genome using Hisat2, and read count using htseq, we performed differential expression using DESeq2 using the default protocol. We used CLASS2 to identify and annotate the alternative splice variants, including previously uncharacterized, of two different long non-coding RNA. One of which is expressed only in SCC, but with variation within samples, which could confirm the heterogeneity of cervical cancer subtypes as proposed by other research groups. The other lncRNA is shown only in ADC samples, which could be a prospect of a biomarker. Further statistical analysis are needed to confirm these primary observations.

Funding: CAPES, CNPQ, FIOCRUZ, FAPESP

”

Exploring lncRNAS in Alzheimer's Disease

Beatriz Miranda, Willian Orlando-Castillo, Silvana Giuliatti

University of Cauca

Abstract

Alzheimer's disease (AD) is a progressive and chronic neurodegenerative disorder, recognized as a multifactorial disease, which limits pharmacological options by the multiple pathways involved in the pathogenesis. To date there are no effective therapies for AD, only the symptoms are treated. Recent studies have shown that non-coding RNAs (ncRNAs), including long-non coding RNAs (lncRNAs), are involved in the pathogenesis of AD. It is believed that there is a strong relationship between the structure of lncRNAs and the functions they carry out. However, due to lack of conservation in its primary nucleotide sequence, functional studies of lNcRNAs with the objective of developing new therapeutic methods imply a challenging process. Since the experimental methods for obtaining tertiary structures are not easy to perform, in silico approaches are valuable tools that contribute to this process of understanding the disease and structural bioinformatics can contribute to the prediction of these structures. Therefore, the goal of this project is to predict the tertiary structures of lncRNAs involved in AD. In this first step, we begin by modeling the structure of lncRNA BACE1-AS, which is highly expressed in AD patients. The unit described by Faghihi et al. (2008) (NR_037803.2) was used. Secondary structure prediction was performed using the Mfold software (Zuke, 2000) and the tertiary structure was modeled using the 3dRNA v2.04 software (Wang and Xiao, 2017). The analysis of tertiary structures was performed using MolProbity (Chen et al., 2010) and Pymol (Delano, 2002) softwares. 29 secondary structures were obtained and the selected model presented a value of -240.61 kcal / mol. BACE1-AS primary and secondary structures were used for the tertiary structure modelling. Five models were successfully generated. After analysis by MolProbity and visual analysis, a model was selected for further studies such as normal mode analysis.

Funding: CNPq

”

Correlation of shared transcriptomic signature between Sickle Cell Disease and Acute Myocardial Infarction patients with Sickle Cell Disease severity

Bidossessi Wilfried Hounkpe, Fernando Ferreira Costa, Erich Vinicius de Paula

Faculty of Medical Sciences, Unicamp

Abstract

While ischemia-reperfusion injury (IRI) is widely recognized as a hallmark of acute myocardial infarction (AMI), it is also a key pathogenic mechanism of sickle cell disease (SCD), raising the question on how shared mechanisms underlie the pathogenesis of these conditions. Analyses of publicly available microarray datasets, integrated with other bioinformatic tools, now allow cross-disease comparisons capable to identify critical components of the pathogenesis of complex diseases. Herein, we aimed to identify a set of differentially expressed (DE) genes in both SCD and AMI, and to use GWAS data to gain further insights on their role in the pathogenesis of IRI, SCD and AMI. Public microarray datasets of SCD with severe phenotype (GSE84632) and AMI (GSE59867) from GEO platform were submitted to meta-analyses using a robust statistical method that allows the identification of genes that were consistently DE in the same direction in both data sets. Functional analysis was performed using FAIME algorithm which scores altered pathways at individual patient level, allowing further selection of most informative pathway by Support vector Machine algorithm. A GWAS catalog was then used to identify risk phenotypes associated with upregulated genes. In order to gain more insights on the importance of identified genes, correlation was computed between DE genes expression and clinical parameters and severity score of another pediatric cohort of SCD. 375 patients were included (80 SCD, 111 AMI, 184 controls). The meta-analyses detected 14 upregulated and 32 downregulated genes. Functional analyses identified pathways related to inflammation and innate immunity, of which 12 classifiers clustered SCD and AMI together. Variants of upregulated genes previously linked to phenotypes in GWAS were potentially associated with vascular inflammation. WASF2, BMP2K and STRADB were positively correlated with SCD severity when GZMK, EIF3M, PIK3IP1 and COX6C showed a negative correlation. Our strategy allowed us to identify a shared transcriptomic signature of SCD and AMI that were correlated with SCD . While observed pathways were consistent with current knowledge on the pathogenesis of IRI, some DE genes had not been previously associated with SCD or AMI, thus warranting additional studies on their role in the pathogenesis and management of these conditions.

Funding: FAPESP grants # 2015/24666-3; CNPq Brazil. grant # 309317/2016

,,,,,,,

Alternative spliced leader trans-splicing patterns among developmental stages of the flatworm Schistosoma mansoni

Daniel Andrade Moreira, Mariana Boroni, André L. M. Reis, Núbia M. G. S. Fernandes, Jéssica S. H. Rios, Sílvia R. C. Dias, Marina M. Mourão, Glória Regina Franco

Universidade Federal de Minas Gerais

Abstract

Spliced leader trans-splicing (SLTS) is an RNA processing mechanism that involves splicing between two distinct RNA molecules. SLTS has been described as a potential RNA regulatory process that occurs in different organisms, including the trematode Schistosoma mansoni, in which at least 46% of cercariae transcripts are processed by SLTS. Here, we performed a deep analysis of SLTS-processed transcripts in five different life stages of S. mansoni, enabling the identification of a large number of transcripts undergoing SLTS and, showing for the first time, stage-dependent alternative SLTS and specific patterns in the parasite. Total RNA of miracidia, sporocysts, schistosomulae and adult worms was isolated and then used to perform cDNA synthesis. SLTS transcripts were enriched prior to sequencing, through a polymerase chain reaction with an S. mansoni SL primer. The SL enriched libraries were sequenced on the Ion Torrent PGM^{TM} System. After quality trimming, only reads identified containing the SL sequence were used on the alignment step. A total of 1, 832, 749 reads were uniquely mapped to 2, 065 different genes in the S. mansoni genome (5th version), grouped within six classes according to the SL acceptor site location (Outron, Outron-1stcodon, Canonical cis-splicing, Exon-middle, Intron-middle or 3'UTR). A chi-square test showed a significant dependence on the parasite developmental stage to the SLTS acceptor site (AS). Moreover, the Outron was the AS most frequently used and SLTS genes within this group undergo SLTS more often in all stages when comparing with genes classified as canonical cis-splicing AS, raising the hypothesis that SLTS in the outron regulates the expression of genes related with general metabolism in S. mansoni, whereas the alternative SLTS seems to assume a stage-specific pattern. In general, there was no correlation between SLTS frequency and gene expression. However, we identified a significant positive correlation between gene expression and the frequency of SLTS when occurring in the middle of adult worm's intron. Indeed, when the SLTS takes place in the middle of introns or exons, the analysis of dinucleotide conservation in acceptor sites showed higher frequencies of weaker AS (others than 'AG'), suggesting that in these cases, SLTS is favored by high gene expression. The classification of SLTS-processed transcripts in functional categories showed ubiquitous distribution with a wide spectrum of functional classes on all stages analyzed. Surprisingly, a chi-square test showed a significant dependence of 3'UTR SLTS transcripts related to transposable elements, and this could be a defense mechanism that regulates the action of these elements. So far, we have shown that the SLTS mechanism in S. mansoni is stage-specific and it is potentially related to other functions beyond those that were previously studied (e.g. resolution of polycistronic transcripts, enhance of transcript stability and translational efficiency). These results may be further extended to other taxa and may contribute to build a more thorough

”’

Transcriptomics approach to identify subtype-specific candidate genes and associated drugs for new therapies in colorectal cancer

Cristóvão Antunes de Lanna, Nicole de Miranda Scherer, Luís Felipe Ribeiro Pinto, Mariana Boroni

Instituto Nacional de Câncer

Abstract

Colorectal cancer (CRC) is the fourth most prevalent carcinoma worldwide, being the third most common in men and the second in women in Brazil. The incidence is related to hereditary factors, eating habits, overweight and physical inactivity. The variety of combinations of these factors results in highly heterogeneous tumors reflecting different prognosis and response to treatment. Different classification strategies have been proposed to characterize tumors more efficiently. The Colorectal Cancer Subtyping Consortium (CRCSC) recently identified four consensus molecular subtypes (CMS1-4) from primary CRC transcriptomic data. Identification of disease-related genes with high potential for drug interactions may assist in the discovery of new targets and more effective therapeutic strategies. This enables repositioning of previously approved drugs to treat other diseases and may reduce the time required to approve new treatments. The aim of this work is to identify candidate genes and associated drugs for the development of new therapies for different molecular subtypes of colorectal cancer from large-scale genomic and transcriptomic data. Gene expression data from 623 patients generated by The Cancer Genome Atlas (TCGA) were used, totaling 623 samples from primary tumor tissue and 51 from tissue adjacent to the tumor. Tumor samples were classified into 4 groups using the CMSClassifier package, with posterior subdivision of CMS4 samples into epithelial and stromal. Unique differentially expressed genes (DEGs) in each CMS subtype were identified with DESeq2 and InteractiVenn. Co-expression modules were constructed using weighted gene correlation network analysis (WGCNA), correlated with subtypes and normal samples, and used in the construction of protein-protein interaction networks using the STRING base. Interactions with low confidence were filtered out and subgroups were identified within each module using the igraph package. Hub genes were selected by the subgroups' connectivity degrees and used to search for drug-gene interactions in the DGIdb database. Molecular-type Drug propositioning were validated using sensitivity data in cell lines from Genomics of Drug Sensitivity in Cancer (GDSC), classified into CMS subtypes from expression data available from the Gene Expression Omnibus (GEO) database using CMScaller, a cell line-specific classifier. Thirty drugs for CMS1, 33 for CMS2 and 33 for CMS4e were identified, 16 of which have already been tested in cell lines. These 16 drugs were evaluated for repositioning in repoDB. Of these, four have not yet undergone cancer clinical trials, and among the others, only two have been tested for CRC, one of them being approved. Seven hubs genes identified within the criteria defined in this work do not have known interactions with drugs. These results demonstrate the potential for the evaluation and implementation of new therapeutic strategies in CRC and the possibility of implementing these analyses in other tumor types.

Funding: Capes, INCA, Ministério da Saúde

,,,,,,

The exon-Junction Complex Proteins MAGOH and MAGOHB are pro-tumorigenic factors in glioblastoma

Fabiana Marcelino Meliso, Wei-Qing Li, André Luiz V. Savio, Bruna R. Correa, Mei Qiao, Pedro A F Galante, Luiz O. Penalva

Hospital Sirio Libanês

Abstract

Glioblastoma multiforme (GBM) is the most aggressive tumor of the central nervous system. In spite of advances in science and medicine, the average life expectancy remains about 18 months after diagnosis and the current treatment are becoming obsolete. Therefore, the search for more effective GBM therapeutic targets is urgently needed. To look for new therapeutics targets, here we investigated, through computation tools and next-generation sequencing data, the role of exon junction complex (EJC) components MAGOH/B, key genes of the post-transcriptional gene regulation mechanism, in low-grade gliomas and GBM. Our results show that a higher expression of MAGOH/B is: i) positively correlated to more aggressive gliomas; ii) significantly related to lower overall survival of GBM patients and iii) with GBM showing worst responses to treatments. We also find that the knockdown of MAGOH/B decreases the GBM cell lines viability and proliferation, but increase their apoptosis. Additionally, we find that MAGOH/B knockdown changes the expression of genes associated with splicing, RNA transport, translation, and cell cycle affected, suggesting an auto-regulation. Interestingly, genes that were alternatively spliced by MAGOH/B KD were linked to RNA stability/processing/metabolism, DNA repair, and stress response, Gene Ontology pathways commonly deregulated in cancer. Furthermore, we have shown MAGOH/B KD reaches RS exons and leads to stop-codon gain and frame change in genes commonly deregulated in GBM. In summary, we believe that MAGOH/B are key genes involved in GBM, which would be investigated as new markers for the disease or novel targets for therapy in the near future.

Funding: PNPD/CAPES, Hospital Sirio Libanês

,,,,,

Using macrophage genes expression to build and validate a molecular model of host-parasite interaction

Felipe Caixeta Moreira, Ana Maria Caetano Faria, Tatiani Uceli Maioli, Leandro Martins de Freitas, Paolo Tieri, Filippo Castiglione

Universidade Federal de Minas Gerais

Abstract

Macrophages are mononuclear phagocytes that constitute the first line of defense against pathogens. They are also the immune effector cells that, upon activation, are able to kill intracellular organisms and are the primary host cells of Leishmania spp. parasites, the obligate intracellular pathogens that cause leishmaniasis. This group of disease has a spectrum of clinical manifestations ranging from self-healing cutaneous ulcers to severe visceral alterations. In mammals, macrophages are the main host for Leishmania amastigote. In regard of this, the aim of this work was to evaluate the genes that are most expressed in macrophages infected with Leishmania. In the developing domain of big data, the role of a data miner is pivotal in the prominent increase in the number of data published The techniques are used in combination with functional transcriptomic, measurement of expression profiles and functional interactions from cells and molecules of many different organisms. The responses of host cells to pathogenic microorganisms are among the most-well studied examples of cellular responses to external stimuli. Pathogen-induced phenotypic changes in host cells are often accompanied by marked changes in transcriptomic expression. In the present study, we collected mining results from results from different databases of cells infected with Leishmania ssp. and we compared those results to differences in THP1 macrophages-gene expression at 12 and 24 hours after Leishmania major infection. We employed a data mining approach, such as the R and Cytoscape software, to filter and select the most prominent genes, we analyses the gene expression profile of 30 THP-1 macrophage diamond genes, 12 and 24 hours after L. major infection. The data were validated with qPCR and we have shown no different expressed genes in THP1 macrophages as we found in the mining data. These analyses provide insights into the interplay between human macrophages and Leishmania parasites, and constitute an important general resource for the study of which genes are the most regulated during the host-parasite interaction. Therefore, there is no difference in gene expression 12hpi and 24hpi, which can be an interesting result that shows the influence of Leishmania major during infection.

Funding: Fapemig, Capes, CNPq, CNR

,

Analysis of long Non-Coding RNAs from RNA-seq Data of Leishmania-Infected Human Macrophages

Flavia Regina Florencio de Athayde, Flavia Lombardi Lopes

FMVA-Unesp

Abstract

Long non-coding RNAs (lncRNAs) are RNAs greater than 200 nucleotides in length, that accomplish a remarkable variety of biological functions. They function as inihibitors or activators of transcription/translation, but with no protein-coding capacity. Macrophages are the primary host cells of Leishmania spp., and constitute a first line of defense against these trypanosomatids responsible for the prevalent zoonotic disease, leishmaniasis. Little is known about the regulatory function of lncRNA in human cells harboring intracelular pathogens. We conducted an analysis using RNA-seq data to identify annotated lncRNAs and alterations in their expression in L. amazonenses and L. major infected macrophages, compared to macrophages exposed to latex beads, as a control for phagocytosis. The main cloud-computing server of Galaxy (usegalaxy.org) was used to align eleven datasets with paired-end reads (GSE-PRJNA290995) to the human genome (version 38) using hierarchical indexing for spliced aligment of transcripts (Hisat2 - Galaxy version 2.1.0). Transcriptome assembly was performed with StringTie (Galaxy version 1.3.4) using annotation Gencode (version 29) to identify transcripts in the data. Next, using StringTie merge (Galaxy version 1.3.4), we created a single assembly GTF file from each group. To characterize their coding potential, we used the software FlExible Extraction of Long non-coding RNAs (FEELnc), and featureCounts (version 1.6.3) was employed to estimate the number of candidate lncRNAs fragments in all paired-end libraries. Abundance of reads were used in differential expression analysis with DESeq2 (R version 3.5), results were filtered to 3107 known lncRNAs, of which 311 were differentially expressed between treatments with FDR-adjusted p-value<0.05 and fold change>2.0. Of 218 differentially expressed lncRNAs in macrophages infected with L. amazonensis versus control, 153 were upregulated and 65 were downregulated. In macrophages infected with L. major, we found 217 differentially expressed lncRNAs, 123 upregulated and 94 downregulated in macrophages, as a result of L. major infection. This study characterizes lncRNA expression signatures in macrophages following infection by Leishmania spp, and suggests a role for non-coding RNAs in immune response to Leishmania infection.

Funding: FMVA - UNESP

”’

Gender-based differences in gene expression and alternative splicing profiles of glioma patients

Gabriela Der Agopian Guardia, Felipe R. C. dos Santos, Luiz O. Penalva, Pedro A F Galante

Children's Cancer Research Institute, UT Health San Antonio, San Antonio, Texas, USA

Abstract

Gliomas represent the most common malignant tumors of the central nervous system, predominantly arising from astrocytes and glial progenitors. The World Health Organization classifies malignant gliomas into 3 histological grades (II-IV) based on the level of malignancy. Glioma standard treatment comprises maximal safe surgical resection followed by radiotherapy and/or chemotherapy depending on the tumor stage. Despite the severe treatment regimen, glioma patients present extremely poor survival rates, e.g., 5-year survival rate of 19.6% for glioblastoma (grade IV) patients. Therefore, the identification of therapeutic targets for the development of more effective treatment approaches is still an active research field. In this context, previous studies have also shown that not only the prevalence of gliomas is higher in men than women, but also men usually have poorer responses to standard treatment. However, the molecular bases of these differences have not yet been fully elucidated from a genomic perspective, thus hampering the development of gender-specific therapeutic options that could substantially improve patients overall survival. In this work, we investigate gender-based differences in gene expression and alternative splicing profiles of primary glioma patients (grades II-IV) to reveal potential modulators of cancer risk and outcome. For each tumor grade, we identify sets of genes which are up-regulated in men or women patients, and whose expression levels do not substantially differ between genders on healthy (cortex) samples. Similarly, we also investigate genes with alternative splicing variants prevalent in men or women patients which do not show significant splicing deregulation on healthy cortex. Next, we show that expression/splicing imbalances are not restricted to protein-coding genes, but also observed in the class of long non-coding RNAs, which corroborates their role on several brain disorders. Finally, we explore the association between the expression levels of all identified genes and glioma patient survival to unveil genes exclusively associated with prognosis of male or female patients. In summary, our study established gender-based transcriptomic differences of distinct malignant glioma grades at both the expression and alternative splicing levels. Altogether, these findings contribute to a better understanding of gliomas aggressiveness and may explain differences in therapeutic responses between men and women patients.

Funding: FAPESP

”

Human Retrocopies and Genetic Expression in Tumor and Normal Tissues

Helena Beatriz da Conceicao, Gabriela Der Agopian Guardia, Pedro A F Galante

IME - USP/ IEP - Sírio Libanês

Abstract

Retrocopies are copies of messenger RNAs reverse transcribed into the genome. Since this duplication process occurs from mature messenger RNAs (without introns), retrocopies are characterized by the conservation of only their parental exons. This characteristic has been used to identify retrocopies since 1980. Also, as a consequence of the lack of promoter regions to allow their expression/transcription, the first works describing retrocopies classified them as "dead on arrival" (i.e., never transcribed). However, nowadays there are a few known mechanisms described in the literature that allow retrocopies to gain expression, such as the obtention of regulatory sequences from neighborhood elements, distant CpG islands or even from cis-regulatory regions. In addition to the label of retropseudogenes, an unexpected number of functional retrocopies have been identified due to the increasing practice of complete genome sequencing and multidisciplinary approaches. Thanks to them, we know nowadays that the mechanism of retrocopy generation has a major role in the genesis of genes in humans and other species. However, this role is limited to a few examples and it is rare the studies that systematically analyze the functionality of retrocopies. For example, in humans, it has been identified approximately 8000 retrocopies, of which only a minority (less than 10%) presents evidence of transcription and functionality. Two important examples are PTENP1 (a retrocopy of the tumor suppressor PTEN) and BRAFP1 (a retrocopy of the proto-oncogene BRAF). In this context, previous works have hypothesized that retrotransposition can be a possible mechanism of neoplasia. It is known that genomic instability can promote retrotransposition activity, with reported cases of pseudogene insertions in tumors. Since many functional retrocopies are still poorly characterized, our objective is to investigate retrocopies with expression in humans, in order to give a global vision of retrogenes contribution to normal and cancer cells phenotype. To explore the possible implications of transcribed retrocopies in the human genome, we used a list of retrocopies detected by an in-house pipeline. Based on this list, we performed an analysis of gene expression using RNA-Seq data from healthy (GTEx) and tumoral (TCGA) datasets. In our results, we gathered general characteristics of retrocopies and their parental genes, such as the number of retrocopies per parental gene and how the retrocopies are distributed in the genome. In healthy tissues, we found that 2/3 of retrocopies show distinct expression patterns among tissues, suggesting some kind of expression regulation. For example, testis has the highest number of expressed retrocopies (4067), while whole blood has only 1767 expressed retrocopies, being the tissue with the lowest number. In tumor samples, we found that 75% of retrocopies are transcribed, with some interesting cases such as three aggressive tumors (STAD, GBM, and AB) having the highest number of transcribed retrocopies (5387, 3853, 4775, respectively), suggesting the possibility of using some of these retrocopies as tumor markers. In the end, our results revealed patterns in terms of retrocopies expression in both normal and cancer that will be further explored, including the investigation of features associated with the selection of retrocopies and the incorporation of Ribo-Seq data to our analyses.

Funding: FAPESP

,,,,,

RNA-Seq of endogenous human stem cells and tumors to identify cancer-specific therapeutic targets

Isabela Pimentel de Almeida, Mainá Bitar, Elizabeth O'Brien, Grace Borchert, Charlotte Woods, Guy BArry

Universidade de São Paulo

Abstract

Stem cells are characterized by their capacity for self-renewal, long-term viability and ability to differentiate into multiple types of specialized cells. Similarly, cancer cells are also capable of self-renewal, which allows aggressive and unlimited tumor growth. Interestingly, pathways that are normally associated with stem cell development overlap significantly with cancer progression. Therefore, endogenous stem cell populations residing outside the tumor are significantly affected by cancer treatments as they target common proliferative signaling pathways. Here we investigate for the first time the similarities and differences between various types of endogenous adult human stem cells and publicly available patient-derived glioblastoma and medulloblastoma primary tumors based on transcript expression via RNA-Seq experiments. Additionally, we profiled the known gene targets of all currently FDA approved drugs for cancer treatment in our stem cells data. The study included Kallisto and STAR/Rsem methods for transcript quantification and also different methods for comparison of expression profiles. Comparing the transcriptomes of normal human stem cells and cancer cells represents an alternative approach to identify better drug targets, with potentially less severe side effects. As proof of principle, we used our data to uncover clinically relevant antisense oligonucleotides (ASOs) targeted to candidate transcripts that were highly expressed in glioblastoma but negligibly expressed in stem cells. We observed a marked decrease in proliferation of primary glioblastoma cell-lines treated with these ASOs. This strategy may be further applied to virtually all cancer types and improve cancer treatment by both assessing existing FDA approved drugs and proposing new targets. Therefore, our findings may support the development of alternative therapies that specifically target the malignant cells within a tumor.

Funding:

,,,,,

Circulating miRNAs can affect the melanoma microenvironment and outcome

Jéssica Gonçalves Vieira da Cruz, Marco Antonio Pretti, Natasha Jorge, Martín Hernán Bonamino, Patricia Abrão Possik, Mariana Boroni

Laboratory of Functional Genomics and Bioinformatics, Oswaldo Cruz Institute (IOC), Oswaldo Cruz Foundation (Fiocruz), Rio de Janeiro, RJ, Brazil.

Abstract

Metastatic melanoma is an aggressive and deadly disease, with high capacity for metastasis and resistance to treatment, resulting in high number of patients dying within 5 years of diagnosis. Various tumors, including melanoma, can interact with their microenvironment and modulate it to enable disease progression, presenting immune evasion characteristics and facilitating metastatic behaviour. To delve the impact of the crosstalk between melanoma cells and tumor microenvironment (TME) on patient's outcome, we accessed RNA-Seq data from 164 metastatic melanoma samples from The Cancer Genome Atlas to characterize their TME using the CIBERSORT. Next, samples were separated into 3 groups by unsupervised hierarchical clustering analysis based on their TME profiles (Jaccard bootstrap mean: G1 = 0.56, G2 = 0.75, G3 = 0.84). The TME profile in each group was distinct, with G1 enriched in naïve, memory and plasma B cells and depleted in resting natural killer (NK) cells, G2 enriched in T CD8 cells, monocytes and M1 macrophages and G3 enriched in M0 macrophages and depleted in plasma cells, T CD8 cells, memory activated T CD4 cells, follicular T helper cells, activated NK cells, monocytes, and resting dendritic cells (p = 0.05, Mann-Whitney test - MW). Overall survival of the groups was compared and G2 patients presented a significantly better prognosis than G3 (p = 0.01, log-rank test, Hazard Ratio (HR) = 0.49, CI.95 = 0.28 0.85). To better understand the interplay between tumor and its TME, we investigated putative interactions between differentially expressed miRNA (miR) and target genes (mRNAs) in G3 compared to G2. We selected miR-targets pairs (MTP) that were predicted in at least one of the databases available in the multimiR package and that were highly negatively correlated to each other(r = -0.4 and p = 0.5), ending up with a list of 139 MTP. We use igraph to represent the network of MTPs, including additional information of gene expression levels, impact on survival, and possible origins of the miRNA. We found interactions that suggested inter- and intracellular regulations, with tumor modulating gene expression on microenvironment cells and vice-versa. One interesting example is the MTP mir-149/NLRC5. The mir-149 is upregulated in G3 and does not impact patients' survival. However, downregulation of its target, NLRC5, has a negative impact on patients overall survival (p = 0.0047, log-rank test, HR = 0.46, CI.95 = 0.27 0.8). NLRC5 is a transcription coactivator that regulates the expression of genes belonging to the antigen presentation pathway. We found many of its targets such as HLA-C, TAP1 and B2M also downregulated in G3, with significant impact on overall survival. Moreover, mutations in the antigen processing and presenting pathway in G3 were associated with increased number of neoepitopes - new and potentially immunogenic peptides generated by mutations (p = 0.05, MW). This can help explain why the immune infiltrate in G3 is so poor on effector cells. Our results point to a crosstalk between melanoma and the TME that can impact the cell types present within the microenvironment and the capacity of the tumor to evade immune surveillance, favouring metastasis and a worse patient's outcome. This knowledge can be used in the future

,,,,,

Characterization of the virome in mosquitoes using a small RNA-based approach

João Paulo Pereira de Almeida, Eric Roberto Guimarães Rocha Aguiar, Roenick Proveti Olmo, Yaovi Mathias Honore Todjro, Jean-Luc Imler, João Trindade Marques

Universidade Federal da Bahia

Abstract

High-throughput sequencing techniques and bioinformatics have guided to the elucidation of viral diversity (virome) in different organisms. Since all viruses produce RNA molecules at some point in their replication cycle, RNA-seq is an efficient and unbiased approach to assess viral diversity. In insects, RNA interference is a major antiviral response and generates small viral RNAs of specific sizes. Virus-derived small RNAs (vsRNAs) produced by the small interfering RNA (siRNA) pathway are phased duplex small RNAs of 20–23nt while products of the piwi-interacting RNA (piRNA) pathway are 24–30nt long and show specific patterns of base enrichment. Different from standard RNA-seq approaches that focus on long RNA fragments, sequencing of small RNAs allows the assembly of viral contigs, but also the inference of antiviral mechanisms and sequence-independent identification of viral contigs based on the pattern of vsRNAs. This strategy overcomes the limitation of sequence similarity searches of highly divergent viral sequences. In this work, we improved our previously developed method to identify viral sequences using small RNA libraries. Our pipeline is integrated into a Perl script requiring a fasta||fastq file as input. After pre-processing, reads are aligned against the host genome, bacterial and transposons reference sequences. Unmapped reads are kept and used to assemble contigs. Viral contigs are then identified by sequence similarity using Blastn or Blastx followed by ORF prediction and protein domain search. Filtered reads are aligned to the assembled contigs to generate a small RNA pattern that is used for hierarchical grouping using UPGMA and Pearson correlation coefficient. The script is run with a single command line in a terminal Linux. Parsed files with viral taxonomy information and plots with read coverage and small RNA profiles per contig are generated as output and are ready to be used in further analysis or publications. We applied our strategy to 56 unpublished small RNA libraries from wild mosquitoes (Aedes, Hemagogus, and Sabethes) collected around the globe. The summary of assembly metrics for viral contigs identified by sequence similarity in all libraries: an average of 43 contigs per library; average N50 of 746nt; contigs size average of 376nt; largest contigs average size of 1968nt. In total, we identified 17 viruses, 9 of which are possibly new viral species. The summary of assembly metrics for contigs without a similar correspondent in the databases: an average of 1866 contigs per library; average N50 of 66; contigs size average of 69nt; largest contigs average size of 465nt. Interestingly, in these last group of contigs, we identified 197 with only siRNA pattern, 62 with only piRNA pattern and 61 with both siRNA and piRNA patterns, all evidence to suggest the origin from unknown viruses of these contigs. Our improved approach allowed us to have an overview of viral diversity and provided information about antiviral defenses of these mosquitoes, knowledge that can be applied in the control and monitoring of virus circulation in natural conditions.

Funding: CAPES, CNPq, FAPEMIG, ZikAlliance, ANR, Labex NET-RNA

,

Neuroblastoma Meta-Analysis for Gene Characterization of INSS Stages

André Luiz Molan, José Luiz Rybarczyk Filho

Instituto de Biociências de Botucatu - UNESP

Abstract

Neuroblastoma is an extracranial solid tumor, very heterogeneous and with highly predictive clinical behavior. It mainly affects individuals under 15 years old and it is classified according to the International Neuroblastoma Staging System (INSS) and the International Neuroblastoma Pathology Classification (INPC). With Next Generation Sequencing (NGS) technologies, performing gene expression profiling of tumors has become more common, especially through RNA-seq. The amount of data generated, however, is large. Thus, the application of meta-analysis and functional enrichment techniques become indispensable for a more effective study. In this paper, based on the RNA-seq gene expression profile of 498 patients, we performed a meta-analysis and a functional enrichment analysis searching for significant gene groups in order to characterize the 5 major tumor stages according to the INSS stage classification. The data were grouped according to these stages and the meta-analysis was performed in the programming environment R with WGCNA package and its function metaAnalysis, which uses the Stouffer method and generates p-values for each of the genes present in the samples. These p-values were corrected by FDR (False Discovery Rate) with the p.adjust function of the Stats R package, generating a q-value. Only q-values lower than 0.05 were considered to be significant. Functional enrichment analysis was done by ADAM R package, comparing, two by two, each of the tumor stages (10 comparisons). For each comparison, genes were regrouped according to their functions based on Gene Ontology (biological processes, molecular functions and cellular components) and pathways from the KEGG repository. For each functional group, q-values were calculated for gene diversity and gene activity. Significant genes obtained with meta-analysis were related to significant groups (only groups with q-value lower than 0.05) obtained by functional enrichment. We found 5163 significant genes in meta-analysis. By relating these genes to the most important functional groups, we noticed an increase in gene and functional specificity proportional to the considered tumor stage. An example of this can be observed when comparing stages 1 and 2 and 1 and 4 for gene activity and biological processes. In the first comparison (1 and 2), we observed 3631 functional groups and 1164 genes. However, in the second (1 and 4) we noticed 189 groups and 330 significant genes related to them.

Funding: CAPES

,,,

Characterizing the global virome of Apis mellifera using a small RNA-based approach

Juliana Armache, João Paulo Pereira de Almeida, João Trindade Marques, Eric Roberto Guimarães Rocha Aguiar

Universidade Federal de Minas Gerais

Abstract

Human activity is increasingly destabilizing ecosystems, causing incalculable and often irreparable damage to biodiversity. In the floristics field, the consequences of human action are already being observed globally, generating a significant loss in native plants and impacting commercial agriculture. One of the factors that contribute to this scenario is the decreasing number of pollinating agents, mainly insects, driven by the increased use of pesticides that are harmful to those animals. Among insects, bees are considered the main pollinators, having a global distribution and attending to a wide spectrum of plants. Besides being sensitive to pesticides, these animals are also susceptible to viral infections, and these two factors combined have led to colony collapses with consequent economic loss. Therefore, knowledge of bee virome can help the development of new strategies to control and prevent viral infections. This study aimed to analyze the collection of circulating viruses in different populations of Apis mellifera using a strategy based on small RNA sequencing data. Twenty-four libraries of A. mellifera small RNAs, originated from South Africa, USA, China, Netherlands and UK were chosen. First, the libraries were pre-processed to remove sequencing adapters and to filter sequences with bad quality (Phred <20). The second step was to remove sequences that mapped to host genome (Apis mellifera) or known bacterial genomes. The remaining reads were subjected to a de novo assembly strategy using Velvet and SPAdes. The resulting assembled contigs were characterized by sequence similarity analysis against NT and NR databases using BLAST. Viral sequences were found in 17 out of 24 libraries, and the results included some viruses known to cause high mortality rates in bee colonies, such as Varroa destructor virus (VDV) and Deformed wing virus (DWV). From all the assembled contigs, some of them showed a significant similarity to known viruses at the nucleotide level, suggesting they are likely new strains of these viruses. In other cases, the similarity to viral sequences was limited to the aminoacid level, which suggests these might be new viral species. Spatial analysis showed that DWV is infecting bees from all over the continents, while other viruses are restricted to some regions. In addition, the contigs that could belong to new viral species were restricted to libraries from Europe. Using this approach we were able to find in our samples viruses that could have both economic and ecological importance, and possibly a new virus that infects Apis mellifera. Unraveling global virome of bees is an important tool to identify and monitor viruses that may cause harm to colonies. This is the first step to help prevent future outbreaks to avoid big economical and biodiversity losses.

Funding: CNPq

,

A machine learning approach to brain region classification

Lissur Azevedo Orsine, Adriano Barbosa da Silva

Laboratório de Biodados, Universidade Federal de Minas Gerais

Abstract

The knowledge of the similarities and particularities of brain regions contributes to understand the hemisphere-specific and holistic organ's biology. The brain can be studied based on different criteria, such as: electrical activity, connectivity, cell types and evolutionary origin. The emergence of RNA-Seq made also possible to characterize the brain according to patterns of gene expression. In this context, we asked ourselves the following question: is it possible to discriminate brain regions based on gene expression? To try to answer this question, we applied a machine learning methodology to the transcriptome of different brain regions. RNA-Seq experiments of two distinct brains containing 22, 318 genes over 10 distinct brain regions were downloaded from the Allen Brain Atlas. The anatomical structures were: frontal lobe (FL), parietal lobe (PL), occipital lobe (OL), temporal lobe (TL), insula (Ins), cingulate gyrus (CgG), parahippocampal gyrus (PHG), striatum (Str), globus pallidus (GP) and cerebellar cortex (CbCx). For the classification, eight machine learning algorithms were applied against the test dataset: Logistic Regression (LR), K-Neighbor Classifier (KNN), Gaussian NB (NB), SVC, Linear SVC (LSVC), Random Forest Classifier (RFC), Decision Tree Regressor (DTR) and Gradient Increase Classifier (GBRT). The best performance algorithm was selected according to 3-fold cross-validation training accuracy score and therefore used in the subsequent analyses. The most satisfactory performance classifier was LR with accuracy of 0.6 (against 0.4 from KNN, 0.5 from NB, 0.1 from SVC, 0.5 from SVC, 0.5 from LSVC , 0.5 from RFC, 0.1 from DTR, and 0.2 from GBRT). The prediction results of the confusion matrix showed that Str was correctly classified in 100% of cases, while OL in approximately 90% of cases, and FL, PL and TL around 30% of cases. It was also possible to observe that PL and TL present a high percentage of reciprocal exchange: around 20% of PL samples were predicted to belong to TL, and approximately 41% of TL samples were classified as PL. Looking at the F1 score, the best ranked brain region was again Str (F1 score = 1.0), followed by FL (0.53), OL (0.42), PL (0.40) and TL (0.30). Finally, the area under the ROC curve ranged from 0.89 (for PL) to 1.0 (for Str). As perspectives, we plan to run the same pipeline for the higher anatomical levels, as well as to compare gene lists per brain region found through the machine learning approach with those obtained from other methodologies.

Funding: CAPES Biologia Computacional

”

Long non-coding RNAs potentially involved with Schistosoma japonicum sexual maturation

Lucas Ferreira Maciel, David Abraham Morales Vicente, Sergio Verjovski-Almeida

University of São Paulo

Abstract

Schistosoma japonicum is a flatworm which causes schistosomiasis, a neglected tropical disease. There is only one efficient drug for treatment, which may lead to resistance emergence. Due to the importance of sexual maturation for the parasite lifecycle and host immunopathogenesis, Wang et al. (Nature Communications 8: 14693, 2017) performed RNA-seq analyses of females and males obtained from 14 up to 28 days post-infection (dpi) in mouse in order to better understand the molecular mechanisms of sexual maturation. They identified protein-coding (PC) genes and specific pathways whose expression levels are related to sexual development; however, this work did not include an analysis of long non-coding RNAs (lncRNAs), transcripts that in mammals were shown to be key regulators of vital processes. There is one paper in the literature reporting the presence of 3, 000 lncRNAs in S. japonicum, but the annotation was performed with an old version of the genome, and only one male and one female RNA-Seq library were used. Our group has recently shown that lncRNAs expression is stage-specific in S. mansoni. Therefore, the aim of the present work is to identify and annotate a more complete set of lncRNAs that complements the most updated PC transcriptome annotation by re-analyzing all RNA-seq datasets in the public domain, including those generated by Wang et al. (2017), to identify stage-specific lncRNAs related to sexual maturation. For this purpose, 66 RNA-seq libraries from five different life-cycle stages were downloaded from the SRA-NCBI. Reads quality control was performed using fastp and aligned against the genome ASM636876v1 using STAR. Uniquely mapped reads were then used for transcripts reconstruction with Scallop, followed by TACO meta-assembly. Coding potential calculation was performed with FEELnc and CPC2. Transcripts classified as lncRNAs were then submitted to annotation with eggnog-mapper to remove possible remaining mRNAs. Synteny analysis was performed between S. japonicum and S. mansoni genomes. The lncRNAs found were included in the transcriptome dataset and expression quantified with RSEM. Weighted gene co-expression network analyses (WGCNA) were then performed in order to identify modules related to sexual maturation. Our pipeline was able to identify 12, 291 lncRNAs in S. japonicum genome. Synteny analysis identified that 80% of all intergenic lncRNAs were contained inside syntenic blocks of at least 5 pairs of orthologous PC genes. WGCNA analysis identified 7 different modules that demonstrate that lncRNAs have a dynamic expression throughout sexual maturation.

Funding: This work was supported by the Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) grant number 2018/23693-5 to SV-A. LM received FAPESP fellowships grant number 2018/19591-2 and DMV received a fellowship from Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). SV-A laboratory was also supported by institutional funds from Fundação Butantan and received an established investigator fellowship award from CNPq, Brasil.

,,,,,,,,,,

NEOANTINGENS, T AND B CELLS IN SQUAMOUS ESOPHAGEAL CANCER

Luciana Rodrigues Carvalho Barros, Paulo Thiago Santos, Marco Antonio Pretti, Nicole de Miranda Scherer, Ivanir Martins, Davy Rapozo, Priscila Valverde, Tatiana Almeida Simão, Sheila Coelho Soares Lima, Mariana Boroni, Luís Felipe Ribeiro Pinto, Martin Hernan Bonamino

Instituto Nacional de Câncer

Abstract

Esophageal cancer (EC) is one of the ten most incident and lethal neoplasms worldwide. The chemotherapy of choice still involves taxane and platinum-based regimens, without any molecular targets. Therefore, it is of utmost importance to better characterize these tumors in order to develop biomarkers and new therapeutical strategies. Esophageal squamous cell carcinoma (ESCA) exhibits high intratumoral molecular heterogeneity that might favor immunotherapy, such as the immune checkpoints blockade. Nonetheless, the success of such therapies depends on the immune-based microenvironment characteristics of the tumor. RNA-seq analysis from 14 tumor and adjacent normal tissue samples from ESCA patients without previous treatment (INCA - CEP 116/11) was performed using Illumina Hi-Seq 2000. Mutations were detected following GATK best practices protocol. Differential gene expression was calculated by RSEM followed by DESeq R package. Class I and II HLA alleles and expression were defined by Optitype and Seq2HLA. ANNOVAR and VEP were used for annotation and gene to protein conversion. NetMHCpan v4.0 was applied to in silico of HLA affinity. Peptides with binding values higher than 500 nM were considered neoantigens. TCR and BCR repertoire were evaluated by MiXCR and tcR R package. Deconvolution analysis of immune subpopulations were performed by CIBERSORT and xCell. TCGA-ESCA samples (n=75) were used as an independent cohort of patients. R software was used for graphic and statistical analysis along with in-house perl scripts. A high number of mutation derived neoantigens and tumor aberrant antigens (TAA) number varied across tumor samples. Although not detected by RNA-seq, four proteins were expressed by the tumor and surrounding areas. All tumors are enriched with immune checkpoint and activators genes compared to normal counterparts, but their expression varied between patients explaining partially why immune checkpoint blockade therapy are not effective against this tumor. Our analysis evidenced a complex immune landscape in ESCA with major macrophages and T cells infiltration. The high number of B cell clones (mostly IgG) infiltrating the tumors and the high active B cell meta-signatures found suggest B cells may play a role in ESCA progression. Also, B cells are significantly correlated with better overall survival and were found in tertiary lymphoid-like structures within the tumor.

Funding:

”’

lncRNAs modulated in response to metformin treatment

Lucio Rezende Queiroz, Izabela Mamede Costa Andrade da Conceição, Marcelo Rizzatti Luizon, Glória Regina Franco

Universidade Federal de Minas Gerais

Abstract

Metformin is among the most widely prescribed drugs. It is used as first-line therapy for type 2 diabetes (T2D) and prescribed for numerous other diseases including cancer. Despite recent advances, the molecular mechanisms underlying metformin action are not fully understood and the role of long noncoding RNAs (lncRNAs) are yet to be associated with metformin response. Using high throughput RNA sequencing, we have analyzed the transcriptional diversity associated with the metformin treatment in human liver cells and identified in a transcriptomic-wide manner hundreds of differentially expressed genes affected by metformin, including lncRNAs. Co-expression networks and guilt-by-association functional analysis allowed us to identify several novel relations between lncRNAs and genes with known function associated with metformin response. These include NEAT1, a lncRNA never before associated with gluconeogenesis repression upon metformin response and a potential regulator to known genes that are associated with noncoding RNA metabolic processes, RNA processing, cytoskeleton organization, transcription activation, and ATP binding. Moreover, network analyses showed that SPATA41, MIR222HG, LINC02348, DUBR, LINC00324, and MIR122HG, collectively those lncRNAs are related to the regulation of transcription, response to nutrients, acute inflammatory response and are potentially attached to already known and novel pathways in the antidiabetic and anticancer effects of metformin. These results suggest that several pathways are regulated by lncRNAs in response to metformin. Our work opens new perspectives on the mechanisms by which lncRNAs that are activated due to metformin treatment are involved in the regulation and control of molecular pathophysiological mechanisms altered in the diseases where metformin is prescribed and thus provides novel candidates for T2D and other diseases treatment.

Funding: CAPES – Coordenação de Aperfeiçoamento de Pessoal de Nível Superior; CNPq – Conselho Nacional de Desenvolvimento Científico e Tecnológico

,,,

Nitrogen catabolite repression in Trichophyton rubrum during the adaptive process to host molecules

Maíra Pompeu Martins, Pablo R. Sanches, Nilce M. Martinez-Rossi, Antonio Rossi

Abstract

The environmental challenges imposed to fungal pathogens require a dynamic metabolic modulation, resulting in the activation or repression of critical factors, enabling the establishment and perpetuation of the infection in their host. Wherefore, to conquer the different host microenvironments, pathogens not only depend on virulence factors but also on metabolic flexibility, dynamically responding to stressing conditions in the host. The dermatophyte Trichophyton rubrum is a pathogenic fungus adapted to degrade keratinized tissues as the skin and nails, utilizing the resulting amino acids and peptides as the final carbon sources. During infection, the proteolytic activity is associated with deamination reactions, resulting in the accumulation of nitrogen mainly as ammonium ions, actively secreted by the fungus through intensive ammonia production. The secreted ammonia raises the extracellular pH, resulting in an alkalinized ambient, acting as a supporting factor to pathogenicity. We evaluated T. rubrum interaction with keratin, in a metabolic perspective, providing information about gene modulation of the dermatophyte during early infection stage after shifting from glucose- to keratin-containing culture media, in comparison to glucose as the carbon source. Analyzing T. rubrum transcriptome using high-throughput RNA-sequencing (RNA-seq) technology, we identified the repression of essential genes related to nitrogen metabolism, as the ammonium transporter MepA, a glutamate synthase, a proline-specific permease, and a urease. These results suggest the activation of an alternative pathway for nitrogen assimilation in the tested conditions, necessary for the fungus survival in the host. The gene expression profiling of the host-pathogen interaction highlights candidate genes involved in fungal adaptation and survival, elucidating the machinery required to the establishment of the initial stages of the infection.

Funding: FAPESP, CNPq, CAPES, and FAEPA

,,,

The importance of long genes in the gene expression of cells affected by Cockayne syndrome

Maira Rodrigues de Camargo Neves, Livia Luz Souza Nascimento, Alexandre Teixeira Vessoni, Carlos Frederico Martins Menck

Department of Microbiology, Institute of Biomedical Sciences, University of São Paulo, São Paulo/SP, Brazil

Abstract

Proteins CSA and CSB play important roles in the transcription coupled repair sub pathway of a major DNA repair pathway, nucleotide excision repair (TC-NER). These proteins are described as responsible for recognizing RNA polymerase II blocked at a bulky DNA lesion. The absence of one these proteins give rise to a monogenic recessive disease called Cockayne Syndrome (CS). The symptoms include premature ageing and central nervous system degeneration, especially affecting both brain development and neurodegeneration. The specific progression of the symptoms of this disease are not easily explained by the cellular phenotype, which affects active genes in the genome of the cell after DNA damage. Even more paradoxically, TC-NER is associated with the removal of bulky lesions in the transcribed strand of expressed genes, which are caused mainly by UV light, and the most affected tissues in CS patients are not exposed to UV light. On the other hand, the nervous tissues are known to have increased levels of mitochondrial activity and are likely to have high levels of endogenous oxidative agents as a by-product of that. The present work investigates the effects of oxidatively generated lesions in the transcription of CSB deficient cells, both induced pluripotent stem cells (iPSC) and neural progenitor cells (NPC), using RNAseq data. We have used potassium bromate (KBrO3) as a source of oxidative stress in order to induce DNA damage by base oxidation. We have identified 3189 differentially expressed (DE) genes in CS NPC after oxidative stress, while only 3 DE genes were found in control NPC, showing that CS cells are much more sensitive to oxidative stress than control cells. The same was observed in iPSC, but not as prominently: 109 DE genes in CS iPSC, while only 1 DE gene was found in control iPSC. We found an enrichment of longer genes in the population of the most downregulated genes in CS iPSC, CS NPC and control NPC, but not in control iPSC, indicating that longer genes in NPC are more affected by oxidative stress. This result supports the notion that DNA damage caused by oxidative stress could be reducing the efficiency of the transcription, making this effect more visible in longer genes which are subjected to random DNA damage by oxidation. Some events of alternative splicing have been observed in CS iPSC after oxidative stress, but not in control iPSC. Principal component analysis (PCA) revealed more variance among long genes both in NPC and in iPSC, pointing to a higher expression dysregulation of longer genes following oxidative stress in CS cells. Interestingly, PCA revealed more variance between samples before and after oxidative stress rather than between control and CS cells for NPC, but not for iPSC. This indicates that the differences in the expression of NPC cells after oxidative stress are more pronounced than in other cells such as iPSC.

Funding: Fapesp, CNPq, CAPES. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

’’’’’

Role of DIMBOA in the fall-armyworm strain diversification inferred with transcriptome differential co-expression

Karina Lucas da Silva Brandão, Natalia Faraj Murad, Aline Peruchi, Celso Omoto, Antonio Figueira, Marcelo Mendes Brandão

Unicamp

Abstract

Which factors induce speciation in insect lineages? That is undoubtedly one of the most fruitful topics of study, with branches that encompass since the origin of biodiversity itself as far as agriculture production and pest control. Secondary compounds from plants have a major role on this process. Benzoxazinoids (BXDs), or hydroxamic acids, are one of the main secondary compounds found in several cereals, providing plant defense against herbivorous insects and pathogens. The main BXD found in corn is the toxic aglucone DIMBOA (2, 4-dihydroxy-7-methoxy-1, 4-benzoxazin-3-one), present also in wheat and rye, although it is absent in rice. DIMBOA elicites variable responses from noctuid larvae of different species of Spodoptera, including increased consumption and growth rates, and antifeedant effect. On S. frugiperda, the fall-armyworm (FAW), however, DIMBOA acts as a feeding stimulant, and improves larvae growth at low concentration. The FAW is a widespread polyphagous moth and the principal pest of corn in South America. The species is distinguished into two host plant-related strains, the corn strain (CS) that feeds preferentially on corn, and the rice strain (RS), that is found generally on rice. To test DIMBOA role in the lineage differentiation of S. frugiperda we reared both CS and RS larvae in artificial diet, either enriched with DIMBOA or without this compound, and evaluated differential co-expression in the transcriptome of both midgut and fat body larval tissues. This approach can point to genes related to DIMBOA response in the FAW and help to clarify its role in the strain differentiation. The 10-day-old larvae of RS reared on experimental DIMBOA-enriched diet grew less than CS larvae under the same condition. Enriched peptidases, reductases and transferases, all them involved in DIMBOA detoxification, showed differences in expression between the two strains. The higher expression of glucosyltransferases in RS larvae midgut compared to CS larvae may be related to a higher activity of glucosylation to detoxify DIMBOA. Glucosyltransferases present in the CS larvae can be more specific than the ones present in the RS, and for this reason CS larvae are more efficient than RS ones to detoxify DIMBOA, what is translated in higher larval growing of the former.

Funding: Fapesp (grants #2012/16266-7; 2011/00417-3)

,,,,

The CD90/Thy1 in Triple Negative Breast Cancer: associations by bioinformatics between dysregulated genes and Signaling Pathways

Marco Lázaro de Sousa Batista, Aline Ramos Maia Lobba, Mari Cleide Sogayar, Ana Claudia Oliveira Carreira, Milton Yutaka Nishiyama Junior

NUCEL (Cell and Molecular Therapy Center), Biochemistry Department, Chemistry Institute, University of São Paulo, Brazil.

Abstract

Breast cancer is the most frequently diagnosed type of cancer among women in the World, with the ductal-invasive triple-negative (TNBC) type being the most aggressive and lethal. We previously demonstrated that the CD90 is a promising diagnostic and therapeutic target for TNBC patients, in the TNBC-derived cell line non-tumorigenic MCF10A overexpressing CD90 (MCF10A/CD90+) and the tumorigenic Hs578T knockdown CD90 (Hs578T/shCD90). The approaches for high-throughput gene expression analysis, such as DEGs identification, and Pathway Enrichment Analysis are well established. Therefore, the aim of this work is to elucidate the possible crosstalk between CD90 and signaling pathways in CD90 TNBC-derived cell line models, through the RNA-seq, applying and developing new System Biology approaches integrating the in house curated TNBC canonical pathways and genes. The edgeR approach has been used for DEGs identification, and EnrichR software, for the identification and comparison of enriched pathways based on KEGG and Reactome databases. A pipeline has been developed to integrate the TNBC biological knowledge, with the DEGs, the Enriched pathways and with the gene set enrichment analysis. A new proposed strategy improved the integration between the DEGs with the Enriched pathways, separating into three DEGs sets: All, Up- and Down-regulated genes. The enrichment analyses was conducted with the sets in MCF10A/CD90+ (718 DEGs) and Hs578T/shCD90 (16 DEGs). Therefore, a new strategy for RNA-seq analysis has been proposed based on the cell models assays, and the new curated Gene sets from Molecular Signature Databases (MsigDB), integrated with the DEGs and enriched pathways. Each assay was conducted using all gene expression profiles, in each curated gene sets on GSEA (Gene Set Enrichment Analysis) method, with fGSEA software. The PCA analysis was adopted to compare the gene and pathways merged between them, using a strategy to calculate the average of each gene in a specific pathway. The weighted gene co-expression network, using the WGCNA software has been used to integrate the enriched pathways and PCA analysis. . Based on the literature, we have identified the TNBC canonical genes (151) and pathways (65), which have been used as a validation set for the identified DEGs and enriched pathways in the first approach. From the MCF10A/CD90+ DEGs, has been found 14 canonical genes, while for Hs578T/shCD90 we have found 3. The PCA analysis has shown two groups of pathways set, segregating MCF10A/CD90+ and Hs578T/shCD90, showing 5 important pathways involved in opposite biological process in these lineages and they will be studied more carefully.

Funding: Support: FAPESP, PIBITI/CNPq, CAPES, BNDES, FINEP, MS-DECIT, MCTI.

,,,,,,

Alterations in whole blood long non-coding RNA expression following Chikungunya viral infection

Maria Fernanda Silva Lopes, Juliana de Souza Felix, Flavia Regina Florencio de Athayde, Mariana Cordeiro Almeida, Nayra Cristina Herreira do Valle, Natália Francisco Scaramele, Flavia Lombardi Lopes

FMVA-Unesp

Abstract

Chikungunya fever is an arboviral infection caused by the Chikungunya virus (CHIKV) and transmitted by mosquitoes of Aedes genus, mainly Aedes aegypti. Mosquito-borne arboviral infections, such as Chikungunya fever, are major challenges for public health and simultaneous circulation of CHIKV and dengue virus makes differential diagnosis difficult. Disease is characterized by rapid onset of fever, severe arthralgia (which can persist for months or years), myalgia, headache and rash, and even neurological manifestations in pediatric cases, from febrile seizures to meningeal syndrome, acute encephalopathy and encephalitis. Gene transcription and translation are intensely controlled by epigenetic processes, ensuring a temporal and tissue regulation of gene expression. Long non-coding RNAs (lncRNAs) can act as an important epigenetic regulator in various cellular functions, such as recruitment of transcriptional regulators, stabilizing mRNAs through protein recruitment, preventing degradation, and others. The aim of this study was to identify differentially expressed lncRNAs between acute and convalescent phases of the disease, on available RNA-Seq data from 16 whole blood samples of pediatric patients, naturally infected by CHIKV (GSE99992). Sequencing reads were aligned with the aid of the HISAT2 tool, using assembly Homo sapiens (b38) hg38 as a reference genome. For lncRNAs identification we employed a pipeline implemented on the Flexible Extraction of Long non-coding RNAs (FEELnc) platform. For lncRNA counting and differential expression analysis, HTseq-count and DESeq2 tools, available on Galaxy platform, were used, respectively. We found 48 differentially expressed known lncRNAs (FDR<0.05), of which 39 were upregulated and 9 were downregulated during the convalescent phase, when compared to the acute phase. LncRNA2Target v2.0 platform indicated that WFDC21P (ENSG00000261040) lncRNA, also known as lnc-DC, was shown to control 16 target genes, associated with dendritic cell differentiation. Thus, at this point we can infer that lncRNA expression is regulated by CHIKV infection in whole blood cells and may affect immune response regulation in pediatric patients.

Funding: CNPq/PIBIC, CAPES (MS and PhD scholarships) and FAPESP (IC and MS scholarships).

,,,,,,,,,

Unearthing Agave secrets: transcriptome analysis of three species suitable for bioenergy production in semiarid regions

Marina Pupke Marone, Fabio Trigo Raya, Lucas Miguel de Carvalho, Maiki Soares de Paula, Sarita Rabelo, Luciano Freschi, Odilon Reny Ribeiro Ferreira da Silva, Piotr Andrzej Mieczkowski, Gonçalo Amarante Guimarães Pereira, Marcelo Falsarella Carazzolle

USP

Abstract

Agaves are plants that present high productivity in dry areas because of their efficient drought resistance mechanisms. Some Agave species are used to produce alcoholic beverages and others to produce fibers, although studies have suggested that some species could be used as feedstock to produce bioenergy in marginal lands. It is the case of Agave sisalana, which is used in the production of sisal fibers; Brazil is the main sisal fiber producer and exporter in the world. This process utilizes only 4% of the leaves; currently, the rest of the material is discarded, but could be used for the production of second generation (2G) ethanol because of its high content of cellulose and hemicellulose, among other compounds. In addition to that, Agaves present low lignin content, a compound that affects negatively the hydrolysis process (recalcitrance) necessary to the 2G production. There is no genome available for any Agave species nor many published studies about the genetic features of the species cultivated in Brazil. In this context, transcriptomic analysis is an appropriate strategy to explore the genetics of these plants without a reference genome. We have sequenced the leaf, stem and root transcriptomes of three of the most fiber-producing genotypes (A. fourcroydes, A. sisalana and 11648 hybrid) collected from a germplasm bank located on a region with low rainfall. We assembled the 3 transcriptomes separately, finding 26, 779, 30, 962 and 28, 320 genes for A. fourcroydes, A. sisalana and 11648 hybrid, respectively. The subsequent analyses were used to shed light on the cell wall complexity and abiotic stress mechanisms utilized by these three species. Altogether with the differential expression and orthologous analysis, we could observe that although all three species have many features in common, their differences come from the expression level. For example, while A. fourcroydes stem presents higher lignification, it possesses a higher expression of the COMT gene, part of the S lignin synthesis pathway, a type of lignin that is less recalcitrant; A. sisalana and 11648 hybrid, on the other hand, have a lower expression of COMT, presenting higher recalcitrance. This suggests that some species have lower S lignin content than others and could be more suitable for 2G production. By analyzing the most highly expressed genes and tissue-specific ones, we have found that the majority of them were related to high temperature and drought resistance. All have higher expressions of heat shock proteins, ubiquitins and LEA, a gene related to hydric stress. Again, the three species present similar strategies, yet some expression level differences exist. Also, a comparative genomics analysis between these Agave species and high and low biomass production species showed us that Agaves have many expanded heat shock protein families, confirming that this is an important mechanism to dealing with high temperatures. Besides, there are four exclusive families related

,,,,,,

Differences in long non-coding RNA expression in Localized Cutaneous and Mucosal Leishmaniasis

Natália Francisco Scaramele, Mariana Cordeiro Almeida, Maria Fernanda Silva Lopes, Flavia Regina Florencio de Athayde, Juliana de Souza Felix, Nayra Cristina Herreira do Valle, Flavia Lombardi Lopes

FMVA-Unesp

Abstract

Leishmaniasis is a disease caused by a protozoa of the Leishmania genus, and it is considered a serious public health problem, being endemic in 98 countries. American tegumentary leishmaniasis (ATL) is responsible for Localized Cutaneous Leishmaniasis (LCL), simplest form of disease, and for more serious clinical evolution forms such as Mucosal Leishmaniasis (MCL). A set of processes called epigenetic mechanisms guarantee time-tissue regulation of gene expression, and one such process is mediated by long non-coding RNAs (lncRNAs) that act in several cellular mechanisms, such as gene expression regulation, leading to gene silencing or activation, recruitment of transcriptional factors, among others. The present study aimed to identify annotated lncRNAs expressed in patients with primary cutaneous leishmaniasis (LCL form) and to compare with those expressed in mucosal lesions (MCL form). RNA-seq data was obtained from NCBI GEO - Datasets GSE3360, consisting of 10 human samples of primary skin ulcers of Localized Cutaneous Leishmaniasis - LCL (n=5) and Mucosal Leishmaniasis – MCL (n=5). Briefly, reads were aligned using the HISAT2 tool taking as a reference genome the assembly Homo Sapiens (b38) hg38. LncRNA identification and annotation were performed through the pipeline implemented in the Flexible Extraction of Long non-coding RNAs (FEELnc) platform. To quantify gene expression and analyze the differential expression of lncRNAs, the HTseq-count and DESeq2 tools were used, respectively, available on the Galaxy platform. Sixteen differentially expressed lncRNAs were found between LCL versus MCL (FDR-adjusted $p<0.05$). Of those, 5 showed increased expression in Mucosal Leishmaniasis, and 11 were more expressed in Localized Cutaneous Leishmaniasis. Using lncRNADisease v2.0 platform we found 3 mRNA (TPST2, CRYBB1 AND CRYBA4) as targets of the lncRNA MIAT, which also function as a sponge for miR-24, and was upregulated in the MCL form. Thus far, our data suggests a change in lncRNA expression profile between different clinical forms (Localized Cutaneous and Mucosal) of the disease, indicating a possible role for these non-coding RNAs in tegumentary leishmaniasis.

Funding: CNPq/PIBIC, CAPES (MS and PhD scholarships) and FAPESP (IC and MS scholarships).

,,,,,,,

Impact of differentially alternative spliced transcripts on proteome of mice infected with different strains of Trypanosoma cruzi

Nayara Toledo, Raphael Tavares da Silva, Tiago Bruno Rezende de Castro, Carlos Renato Machado, Andréa Mara Macedo, Mariana Fioramonte, Daniel Martins de Souza, Glória Regina Franco

Universidade Federal de Minas Gerais

Abstract

Since the description of Chagas disease, caused by the protozoan parasite Trypanosoma cruzi, the reasons for the different clinical manifestations of the disease in humans have yet to be completely revealed. Our group has previously shown that different strains of T. cruzi (JG- T.cruzi II and Col1.7G2-T. cruzi I) had a differential tissue tropism in BALB/c mice upon infection. Evidences that the genetic background of different mice lineages is contributing for changes in the differential tissue distribution of T. cruzi during infection were also found. Studies on differential gene expression, aiming at elucidating which host genes could be modulated by distinct parasite strains, were conducted using JG, Col1.7G2 or an equivalent mixture of both strains. Alternative splicing is a regulatory mechanism of gene expression in which different exons and introns of the same pre-mRNA may be skipped or retained to produce distinct mature mRNAs. In recent years, this mechanism has been shown to be a major source of cell-specific proteomic variation in mammalians. Thus, the purpose of the present study is to integrate mass spectrometry-derived proteomic data from BALB/c infected hearts with the same T. cruzi strains and the above-mentioned RNA-Seq data. We performed quantification of gene expression at the transcript level and its all distinct isoforms. Comparing Col1.7G2 infected mice with control group, we identified a total of 594 differentially expressed transcripts with false discovery rate less than 0.05, including 543 upregulated and 51 downregulated. Comparing JG with control group, a total of 901 transcripts were considered differentially expressed, including 256 upregulated and 645 downregulated. Functional enrichment analysis showed Col1.7G2 induced a higher inflammatory response while JG exhibit a weaker activation of immune response genes. Furthermore, JG-infected mice showed a notable reduction in expression of genes responsible for energetic metabolism, mitochondrial oxidative phosphorylation, and protein synthesis. Splicing events were frequent, including an increase in the number of skipped exons, increase in introns retention and increased in usage of alternate 5′ and 3′ splice sites. Our future steps include to correlate of these results with proteoforms identified by mass spectrometry.

Funding: FAPEMIG, CAPES

”

CodAn: predictive models for the characterization of mRNA Transcripts

Pedro Gabriel Nachtigall, Andre Y. Kashiwabara, Alan Durham

Laboratório Especial de Toxinologia Aplicada (LETA), Instituto Butantan, São Paulo, Brazil

Abstract

The complete characterization of the coding sequences (CDSs) and untranslated regions (UTRs) of transcripts is an essential step on transcriptome annotation and expression profile analysis. First, it defines which proteins should be synthesized by the messenger RNAs and are part of the proteome of the organism. The incorrect characterization of CDSs can lead to the prediction of non-existent proteins. Wrong protein predictions can eventually compromise knowledge if annotation databases are populated with similar incorrect predictions made in different genomes. Also, the correct identification of CDSs is important for the characterization of the UTR landscape, whereas the 3'UTR and 5'UTR are known as important regulators of the mRNA fate and translate process. Here, we present CodAn, a new computational approach to predict CDS and UTR sequences directly from transcriptome sequences of any Eukaryote species, such as RNAseq assembly data. CodAn can be applied to full or partial transcripts and presents a better performance predicting the whole CDS than other approaches. CodAn requires low computational resources and can be used on any standard desktop computers, and, for large jobs, can use the parallel processing capabilities of large multi-core servers. The data generated by CodAn can be used to improve genome annotation and help further experiments focused on understanding the evolution and biology of CDS and UTR sequences.

Funding: This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001 - Process number 88887.177457/2018-00.

,

piRNAs expression profiles: Estimates and insights found in four human tumor tissues

Ricardo Piuco, Pedro A F Galante

Abstract

Small RNAs plays several roles regulating gene expression, particularly, PIWI-interacting RNA (piRNA), ranging from approximately 21 to 35 nucleotides act suppressing transposable elements and coding mRNAs. Association of these RNAs with piwi argonauts proteins forming the silencing complex as well as their functional interactions have been discovered in recent years. However, it is still mostly restricted to germlines. Knowing that alterations in this complex are related to increased proliferation and invasion of cancer cells, we seek to improve the quantification of piRNAs and identify expression profiles in somatic and tumor context. First, we obtained data from 2359 small RNA sequencing studies from four tissues: breast, colon, prostate and pancreas (1252, 472, 149 and 486 samples, respectively). Using piRNAdb.org database, we found an average of 170 piRNAs expressed in tumor samples and 130 in normal tissues. Compared to the total of 27700 known human piRNAs, we noticed that only a small part is present, however we observed distinct expression profiles between the analyzed conditions, providing evidence to further refine the evaluated parameters and benefit the entire research field. In parallel, we sought to evaluate predicted targets for these piRNAs that might has an oncogenic role or that could be used as biomarkers. For example, NEFL, target of hsa-piR-28400, has been studied in association with microRNAs and cell proliferation in glioblastoma and breast tumors. Then, to expand the knowledge of piRNA profiles, we grouped the samples according to their tumor size, lymph node and metastasis characteristics. This restriction allowed the observation of piRNAs that could be attributed to each of the situations mentioned, for example, there is a greater amount of piRNAs expressed in T3-classified breast tumor (5 piRNAs) compared to T2 samples (1). Different results from those found when evaluating normal versus tumor samples, where more piRNAs are expressed in healthy tissue, as an example, we found 37 piRNAs more expressed in normal colon tissue compared to tumor (11 piRNAs). At the end of our observations we believe that even when comparing samples according to their broader characteristic we identified distinct piRNA expression profiles and by further investigating the sample classifications and predicted targets information we provide good parameters to be used by the scientific community. Results that will be the basis for further studies and/or refinement to early diagnosis, disease progression or even the development of more accurate treatments for patients of these and other tumors.

Funding:

”

Circular RNAs contribute to tumorigenesis and tumor progression in colorectal cancer

Vanessa Galdeno Freitas, Pedro Alexandre Favoretto Galante, Paula Fontes Asprino

Interunidades em Bioinformática - IME USP / Instituto de Ensino e Pesquisa Sírio Libanês - IEP HSL

Abstract

Circular RNAs (circRNAs) are a new class of RNA that form covalently closed continuous loops. Like long noncoding RNAs, the effective functions of circRNAs mainly depend on the characteristics of their sequences and structures and there is a growing body of functional roles assigned to circRNAs. For example, circRNAs act as miRNA sponges indirectly modulating gene expression of protein-coding genes. Furthermore, dysregulation of circRNAs expression have been associated to several human pathologies, such as cancer. In terms of expression, studies have reported hundreds of circRNAs as more abundant than their corresponding linear mRNAs not only in tissues, but also in the blood. Colorectal cancer (CRC) is the third most commonly diagnosed cancer and the fourth leading cause of cancer-related deaths in the world. Studies in CRC cell lines and CRC tissues show an overall reduction in circRNA abundance compared to healthy tissue, allowing CRC cells unexpected and uncontrolled proliferation, for example. Here, we have characterized and studied the expression profile of circRNA in CRC cell lines aiming to better understand the role of these RNAs in the tumorigenesis and cancer progression. First, RNA sequencing (RNA-Seq) data from two commercial cell lines from the same primary and metastatic CRC patient (SW480 and SW620 respectively) were performed in an Illumina NextSeq platform. Next, two methods of RNA sequencing library preparation were used: i) the standard protocol suggested by Illumina; ii) and an in-house protocol to improve the detection of circRNAs. Finally, all RNA-seq data were aligned to the reference genome (GRCh38) and used as input to identify circRNAs through the CircExplorer algorithm and further in-house computation pipelines. RSEM and Kallisto methodologies were also used to generated the gene expression profile from all cell lines and library preparations. R package EdgeR was used for the normalization and for selecting the differentially expressed genes. First, our results show that our in-house protocol of RNA sequencing library preparation detected 70% more circRNAs than a standard preparation commonly used. Next, we evaluated the number of circRNA expressed in the primary (SW480) and metastatic (SW620) cell lines. We found 4, 024 circRNAs differentially expressed (FDR < 0.05; log2fold change >2 or <-2), 1, 964 up regulated and 2, 060 down regulated circRNAs in primary tumor versus metastatic cell lines. By evaluating the cellular pathways of genes generating circRNAs, we identify several candidates involved in tumorigenesis and cancer progression, such as lysine degradation, EGFR tyrosine kinase inhibitor resistance, RNA transport, proteoglycans in cancer, focal adhesion, regulation of actin cytoskeleton, and adherens junction. In summary, we believe that our work has produced novel and pivotal information to a better understanding of the functional role of circRNAs in origination and progression of colorectal tumors.

Funding: CAPES

’’’’’

The effect of genetic diversity in differential gene expression analyses using RNA-Seq data

Victor Mello, Ana Letycia Basso Garcia, Fernando Henrique Correr, Guilherme Kenichi Hosaka, Amanda Ghelfi Dumit, Gabriel Rodrigues Alves Margarido

ESALQ - USP

Abstract

Differential gene expression studies focus on discovering which genes are more highly expressed in a certain condition, tissue or group of organisms. Statistical tests for differential expression using RNA-Seq data often rely on estimates of both the average expression levels and dispersion (variance) of transcript abundances for the contrasting groups. This is done by sequencing biological replicates subjected to the same experimental conditions. Researchers commonly use clones to compose these groups in many cases where it is convenient, such as in vegetatively propagated plants, in order to obtain the highest possible uniformity among the replicates. However, this approach restrains the generalization of results to only a few genotypes, and those may not be valid in a broader sense. In this work, we compare the outcomes of differential gene expression analysis when using a strategy based on clones (SBC) or on diverse genotypes (SBDG) of sugarcane as biological replicates. The samples consisted of sugarcane top internodes grouped by the soluble solids level (Brix), namely Very Low, Low, High and Very High Brix. Within each group there were three biological replicates, which included clones of the same genotype in one set and three different genotypes in the other. The 24 samples, 12 from each set, were utilized to perform a de novo transcriptome assembly, resulting in 262, 281 putative genes. We found that the common biological coefficient of variation with the SBDG was about twice higher than with the SBC. The total number of differentially expressed genes (DEG) was equal to 28, 699 and 10, 380 in the clone and diverse approaches, respectively, taking into account up and downregulated genes. A functional enrichment analysis showed more Gene Ontology (GO) enriched terms directly related to the contrasting phenotypic traits for the "Very Low Brix against others" contrast (FDR < 0.01) using the SBDG. On the other hand, for the other proposed contrasts, the SBC resulted in many more statistically significant enriched terms, although none of them was related to sugar yield or carbon partitioning. The obtained results suggest that using clones as biological replicates minimize the variance of transcript expression levels, resulting in a higher statistical power to detect DEG and GO enrichment, but both might be not representative of the phenomena of interest. These conclusions highlight the existence of bias depending on the sample and replicate choices, which should ideally present a balance between the expected statistical power and the biological meaning of the results.

Funding: FAPESP

”

Identification of alternative splicing variants that are susceptible to NMD pathway by a bioinformatic approach

Vinicius da Silva Coutinho Parreira, Letícia Graziela Costa Santos de Mattos, Fabio Passetti

Laboratory of Gene Expression Regulation, Carlos Chagas Institute, Fundação Oswaldo Cruz (Fiocruz), Curitiba, PR, Brazil

Abstract

DNA high-throughput sequencing associated with bioinformatics approach, permits to perform in silico analysis of genomes and transcriptomes. The last GENCODE release provide 19, 975 genes related to more than 83, 000 proteins. This difference can be associated with RNA post-transcriptional modifications, mainly associated to alternative splicing. Alternative splicing increases transcriptomic diversity due to alternative recongnition of intron splice sites. Although alternative splicing is strictly regulated, some errors may occur, such as the insertion of a premature termination codon (PTC) in the transcript and such molecules could result in a truncated protein. When a termination codon is located at least 55 nucleotides downstream of a exon-exon junction, it is called PTC and the transcript is molecularly marked for degradation by the nonsense mediated decay pathway (NMD). The computational prediction of NMD events or the manual annotation is used by the Ensembl project to provide organism-based NMD datasets. SpliceProt is a protein sequence repository created by our research group that aims to predict splice variants and perform their in silico translation. We aim to create a subgroup of SpliceProt that will have a PTC identification for transcripts, consequently removing these variants of in silico translation step. We have updated the SpliceProt repository to receive Ensembl transcript information for human. The rat and mouse datasets is ongoing. Using a ternary matrices methodology, which received mapping coordinated produced by the BLAT software, we have identified 222, 462 human splice variants that corresponds to 61, 122 human genes. According to preliminary results using human known control mRNAs with NMD-targets and non-NMD-targets, we were able to correctly identify all NMD events in concordance to the NMD Ensembl annotation (e.g. LENEP and CRYZ). Currently, we are performing the PTC identification in the SpliceProt datasets using the NMDClassifier software. In addition, we will integrate the NMD classifier in the repository's pipeline to enable automate future SpliceProt releases.

Funding: National Council for Scientific and Technological Development (CNPq), Coordination for the Improvement of Higher. Education Personnel - Capes, Carlos Chagas Institute - Fundação Oswaldo Cruz (Fiocruz)

8 — Systems Biology and Networks

””

In silico identification of transcriptional regulatory pathways in Leptospira biflexa biofilms

Artur Filipe Cancio Ramos dos Santos, Mariana Teixeira Dornelles Parise, Doglas Parise, Paula Carvalhal Lage Von Buettner Ristow, Vasco A de C Azevedo

UFMG

Abstract

Bacteria of the genus Leptospira comprehend 65 genomic species including pathogenic, intermediate and saprophytic groups. Pathogenic leptospires are the etiologic agent of leptospirosis, a disease of public health and veterinary public health impacts worldwide. Biofilms improve survival of microorganisms in hostile environments and are related to various medical conditions. Leptospira form biofilms in vitro and in vivo. Nevertheless, the regulatory mechanisms of biofilm formation in Leptospira are poorly known. Saprophytic Leptospira biflexa shares several genetic and functional similarities with pathogenic species and can be used as a model to study biofilms. In this study, we aimed to identify transcriptional regulators involved in biofilm formation in Leptospira biflexa and to describe the regulatory pathways of these regulators. Firstly, we selected transcriptional regulators predicted for Leptospira biflexa in P2TF database. Secondly, we conducted a similarity search using Protein BLAST with those regulators against previous data from a Leptospira biflexa transcriptome analysis of biofilm versus planktonic cells, in two time points: 48 h (mature biofilm) and 120 h (late biofilm). Transcriptomic data is publicly available under BioProject accession number PRJNA288909. After identifying biofilm regulatory genes in L. biflexa transcriptomic data, we checked for their expression levels to understand if a particular regulator was contributing positively, negatively or being neutral in the context of biofilm regulation. Finally, we performed a functional annotation, in order to classify all the regulators found using COG database. In total, we predicted 138 transcriptional regulators for L. biflexa, comprising sigma factors, two-component systems, response regulators and other DNA-binding proteins. Among those, we identified 38 (27.5%) regulators as participating in the biofilm phenotype. From the results analyzed so far, we found that the sigma factor LEPBI_II0101 integrate a network alongside with other sigma factors, response regulators, RNA-polymerase subunits genes and two-component systems in all transcriptomic comparisons, leading us to infer that this regulator is important to sense environmental changes and modify expression. We also found that the alternative sigma factor FliA positively regulates motility and chemiotaxis, and interacts with other flagellar proteins in the mature biofilm. Motility and chemiotaxis are pointed to be important for biofilm formation in other species. Our results also evidenced the toxin–antitoxin system VapBC, which contributes with RNase activity in the late biofilm. Our work is novel in describing the regulatory mechanisms of Leptospira biofilm formation and will shed light on the intricate regulatory pathways of this phenotype.

Funding: CNPq (UNIVERSAL MCTI/CNPq No 01/2016; Process: 425526/2016-0). This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001

,

A study of AGN inference with Tsallis Entropy

Cassio Henrique dos Santos Amador, Fabrício Martins Lopes

Universidade Tecnológica Federal do Paraná

Abstract

In the field of gene networks, network inference is an open problem. This inference is handicapped by the low number of samples and the great network complexity (number of genes and even more interactions between them). These networks can be modeled using Probabilistic Boolean Networks, where the gene relationships are modeled as Boolean operators, and these operators can depend on one, two or more genes. Besides, the network structure can be inferred if the correct criterion function is chosen, which can be the information entropy, for instance. Previous work has shown that Tsallis entropy as a criterion function is more accurate to infer network structures than traditional Shannon entropy, if one can find the best non-extensive parameter. This parameter is related to the interaction distance between elements, the complexity of this interaction, and the possible absence of probability configurations between the elements, in this case, the genes. The present work investigates artificial gene networks (AGN) to analyze the reasons for this entropy to be more efficient, and what is the effect of a network structure topology on the non-extensive parameter used for its inference. The use of artificial gene networks allow a controlled environment, and the analysis with a larger number of samples than experimental results could allow for. Different genes with various number of links in the network, from different scale-free networks were studied. We also analyzed the relationship between these different genes and the inference of the network as a whole, and the results are compared with inferences obtained from the Shanon entropy.

Funding: UTFPR

””

Global co-expression network analysis unveils important aspects of evolution and transcriptional regulation in soybean (Glycine max)

Fabrício de Almeida Silva, Fabricio Brum Machado, Kanhu Charan Moharana, Rajesh Kumar Gazara, Thiago Venancio

Universidade Estadual do Norte Fluminense Darcy Ribeiro

Abstract

Soybean (Glycine max (L.) Merr.) is one of the most important crops worldwide, representing a significant fraction of Brazilian GNP. Gene co-expression networks (GCN) have been largely used to elucidate regulatory complexity and evolution of genes and their functions. Here, we have reconstructed a GCN using 1298 publicly available samples from 12 distinct tissues. Sequencing reads were mapped against the soybean genome (Wm82.a2.v1) and relative transcript abundance estimated in Transcripts per million mapped reads (TPM). The network was reconstructed and visualized with the R packages WGCNA and igraph, respectively. The top 10% most highly connected genes with the highest module membership were considered intramodular hubs. We explored the network properties and found critically important modules that are up-regulated in specific tissues. Enrichment analyses of these modules revealed biological processes and pathways that are essential to some particular tissues and may have elementary contributions to the plant development. We also identified transcription factors (TFs) among intramodular hubs, which may be important regulators, shaping the transcriptional landscape in particular tissues. The top hubs for each module were identified and we found that they tend to encode proteins with critical roles, such as succinate dehydrogenase, peroxidases, and RNA polymerase subunits. Further, we analyzed the distribution of soybean paralogous genes across the network modules to better comprehend the fate of duplicate genes in polyploid organisms. Most of the duplicate gene pairs were present in different modules, supporting their subfunctionalization in different tissues and providing insights on the evolutionary importance of polyploidization in soybean genome complexity and evolution.

Funding: CAPES, CNPq, FAPERJ

,

Reconstruction of metabolic pathways of Klebsiella spp. bacteria for improve the biologic control of Mediterranean fly (Ceratitis capitata).

Luis Augusto Franco López, César Alberto Bravo Pariente

Universidade Estadual de Santa Cruz, Ilhéus, Bahia, Brazil

Abstract

Pests are known to cause significant damage to crops and affect agriculture productivity. Mediterranian fruit fly (Ceratitis capitata) is a serious horticultural pest, it attacks a range of cultivated fruits and vegetables. Several findings show that bacteria are present in the Mediterranean fruit fly microbiota and one of the most important bacteria is Klebsiella spp. Nowadays, scientists develop multiple methods to control pests as the Sterile Insect Technique (SIT) to release infertile males produced by irradiation. There are some metabolic pathways in Klebsiella spp. that can improve the behavior of the irradiated Mediterranean fruit fly males at the moment of liberation. The possible relationship between these organisms produces multiple quantities of biological data, to explain it, is necessary the use of mathematical models and computational scientific techniques. Processing these data in an efficient way facilitates its interpretation and scientific application. The objective of the present work is to reconstruct metabolic pathways of three strains of Klebsiella spp., from genome scale and choose one strain that can improve the biological control of C. capitata. Two approaches were used; 1) Stoichiometric reconstructions of pathways and 2) Directed graphs to analyze and evaluate the pathway analysis of the metabolic pathway generated with Elementary Modes (EM) analysis. From genome scale the sequences of the three bacteria were used to find possible metabolic pathways metabolites present in the nitrogen metabolism at these sequences using BLAST tool, those genes are related with symbiotic interaction between bacteria Klebsiella spp and C. capitata. The metabolites in the sequence indicated a possible metabolic pathway in the organism. Using KEGG database with the possible metabolic pathways involved we generate a directed graph and with it, then we produce a stoichiometric matrix with rows representing the metabolites that participate in reactions and columns with the number of molecules of metabolites involved (stoichiometric coefficients) in one reaction. Based on this information we analyzed and evaluate the pathway analysis with EM. The above-discussed considerations suggest that we can identify possible metabolic pathways involved in the process of symbiosis between Klebsiella sp. and C. capitata and we can generate a strategie to improve the production of infertile males of C. Capitata for the liberation and population control.

Funding:

,,,,,,

Integrated transcriptomic and metabolomic analyzes applied to cane-energy: a new variety of cane with high biomass productivity

Jovanderson Jackson Barbosa da Silva, Luís Guilherme Furlan de Abreu, Nicholas Vinícius Silva, Antônio Pedro de Castello Branco da Rocha Camargo, Camila P. Cunha, Gonçalo Amarante Guimarães Pereira, Marcelo Falsarella Carazzolle

Universidade Estadual de Campinas

Abstract

Energy-cane (CE) is a commercial hybrid originated from the crossing of the same parents as sugarcane, Saccharum officinarum and S. spontaneum, but has very distinct characteristics, such as a high fiber content (13.5% higher), low sucrose content (4.1% lower) and higher productivity (85% higher) compared to commercial sugarcane hybrids (CA). These characteristics make energy-cane a promising plant for the production of first and second generation bioethanol and bioelectricity. Although sugarcane is widely studied from a genomic and molecular biology point of view, very little knowledge has been generated about energy-cane, in particular, in order to expand the development of these new hybrids through the use of genomics and systems biology. In this context, this study has generated transcriptomics (RNA-Seq) and metabolomics data from various tissues of energy-cane and sugarcane during the night cycle (18 hours and 24 hours) and day cycle (6h and 12h). For transcriptomic data, several bioinformatics pipelines were developed to evaluate the best approach for assembling the energy-cane transcriptome using de novo approach or guided by the reference genome of S. spontaneum (one of the parental species). An analysis of the metabolites profile in leaf and culm tissues during the night and daytime variation was also performed. De novo transcriptome assembly was performed by Trinity software and guided assembly was performed by the combination of Hisat2 and StringTie software. In both cases, the transcripts obtained were quantified by the Kallisto software. Separation into coding and non-coding transcripts was performed by the combination of Transdecoder and RNAsamba software. The metabolic profile analysis was performed on a Q-TOF Ultima-API mass spectrometer, with ESI ionization source, coupled with a UPLC Acquity and processed by the Global Natural Products Social Molecular Networking (GNPS) platform. Although reference-genome-guided assembly generated far more transcripts (25, 065) compared to de novo assembly (14, 825), some transcripts were obtained exclusively by de novo assembly representing candidate transcripts specific to energy-cane. From the analysis in GNPS, 246 metabolites in CE and CA were recovered in 6h, of these only 26 were identified, at 12h, 202 compounds, 18 were identified, at 18h, 230 compounds, of these 22 were identified, at 00h, 220 metabolites, being 19 identified, at 6h (24 hours after the first collection), 129 compounds, 14 identified. The integrated analysis of these omics approaches is a fundamental step to improve our understanding of the molecular biology of energy-cane compared to sugarcane, generating new insights for the development of more productive varieties.

Funding:

”

Overcoming challenges in the metabolic reconstruction process: A promising approach to the MDRAB problem.

Juliana Simas Coutinho Barbosa, Pablo Ivan Pereira Ramos, Marisa Fabiana Nicolás

Laboratório Nacional de Computação Científica

Abstract

Acinetobacter baumannii is a human opportunistic pathogen associated with multidrug resistant phenotypes. Furthermore, this pathogen is responsible for several outbreaks in intensive care units around the globe. For that, the World Health Organization considers A. baumannii as a top priority when it comes to the need for new antibiotics. In this context, genome-scale metabolic reconstructions (GEMs) are promising tools to help understand the mechanisms behind antimicrobial resistance in MultiDrug Resistant Acinetobacter baumannii (MDRAB) and find promising drug targets for new antibiotics and also synergistic effects to antimicrobial drugs already known and used in therapy. However, the reconstruction process has many challenges, little instructions and is very demanding when it comes to time and effort. In this project, we aim to build a comprehensive metabolic model of Acinetobacter baumannii strain ATCC 17978, capable of predicting cellular responses to genetic and environmental disturbances. Meanwhile, we intent to present a transparent methodology that can assist in the reconstruction and curation processes. Moreover, flux balance analysis simulations will be conducted for case studies seeking drug target prioritization and the elucidation of the antimicrobial resistance mechanisms. Initially, we built 3 automatic reconstructions by mapping the organism's proteome to the KEGG and MetaCyc databases and also to a GEM of a closely-related organism iATCC19606, which comprises the metabolism of A. baumannii strain ATCC 19606. Those 3 reconstructions were merged into a comprehensive automatic reconstruction, putting together as many metabolic functions based on genomic evidence as possible, in order to reduce manual curation efforts when filling in eventual gaps. The resulting automatic reconstruction has 2407 reactions and 2918 metabolites. The great number of reactions and metabolites is most likely due to empty reactions and duplicate reactions and metabolites, which will be removed during manual curation, in a process called reconciliation. Once ID reconciliation is performed, that number is expected to decrease significantly. During the curation process, the model's ability to replicate experimental data accurately is closely monitored, until a satisfactory accuracy threshold is reached (around 80%). That step is called validation and it will dictate when the model is ready for predictive simulations, which should lead to valuable insights about the pathogenesis of A. baumannii.

Funding: CAPES

”

A Network-Based Approach to Study lncRNA associated with Posttranscriptional Regulation Pathways in Hepatocites Treated with Anticancer Drugs Through the Use of Outdated Microarray Data

Giordano Bruno Sanches Seco, Agnes Alessandra Sekijima Takeda, José Luiz Rybarczyk Filho

Instituto de Biociências de Botucatu - UNESP

Abstract

The aim of this work is to search for lncRNA probes in "outdated" microarray data to create a lncRNA-PPI (Protein-Protein Interaction) network to study posttranscriptional regulation of 2 anticancer drugs: etoposide and lomustine. Microarray data for both drugs was prospected from the OPEN TG-GATES Project which used Affymetryx chips to measure gene expression in normal hepatocites (control) and drug treated hepatocytes (case) in high, middle and low doses for 3h, 8h and 24h. The raw data was pre-processed in R enviroment with Affy package from bioconductor repository and normalized with robust multi-array average (RMA) method. A list of lncRNA symbols was prospected from HGNC in order to search for lncRNAs in the microarra's probes. In total, 5 lncRNAs were found: Dancr, EGOT, GAS5, MALAT1 and TUG1. For each of the selected lncRNAs, a RBP-lncRNA (RNA binding proteins) network was prospected in the Starbase database. The mRNAs in each of these networks were then used to prospect 5 PPI networks in the STRING database, using 'Database' and 'Experiment' as interaction types and with a score higher than 0.7 in order to avoid false-positive interactions. The 5 RBP-lncRNA and 5 PPi networks were concatenated into a single network and duplicated interactions were removed. The final network was renderized in Cytoscape, where MCODE plugin was used to derive clusters/modules from the network, with degree cutoff of 2, node score cutoff 0.2, K core 2, depth 100 and loops included. The Bingo plugin was used to perform functional enrichment of the whole network and the 12 modules. As expected the network and clusters are highly enriched for GO (Gene Ontology) terms such as mRNA catabolic process, ncRNA metabolic process, posttranscriptional gene silencing by RNA, nuclear mRNA splicing (via spliceossome), etc. We took the ratio between the high-24h cases expression for both drugs and plotted it over the network. Overall both drugs induces upregulation in most transcripts in the network. For example pre-mRNA-splicing factor SYF2, which is highly upregulated in both drugs, is associated with positive regulation cell proliferation and DNA damage checkpoint and both drugs are know to induce apoptosis via DNA damage. More interesting peharps are the different expressions in the networks, for they might point to specific metabolic steps in each drugs mechanism of action. CDK19 is a well know positive regulator of apoptotic pathways and is upregulated in etoposide's network but downregulated in lomustine's.

Funding: CNPq

”’

HOMOLOGY MODELING AND MOLECULAR DOCKING STUDIES OF ARYLALKYLAMINE N-ACETYLTRANSFERASE (aaNAT) of Aedes aegypti

Maria Angélica Bomfim Oliveira, Fabrício Santos Barbosa, Tarcisio Silva Melo, Bruno Silva Andrade

Universidade Estadual do Sudoeste da Bahia, Brazil.

Abstract

According to World Health Organization (WHO) data, about 700, 000 people die each year due to diseases transmitted by the Aedes aegypti mosquito: Dengue, Chikungunya and ZIKA, among other arboviruses. Additionally, this mosquito is capable of reproducing in urban environments, as well as act as vector, transmit and replicate diferent type of virus, and becoming a great problem of public health in many undeveloped countries. One way for controlling this vector is studying its metabolism, and identifying important protein target which can be modeled and used for. The enzyme arylalkylamine N-acetyltransferase (aaNAT) is essential in the process of cuticle sclerotization and mosquito development. Therefore, the aim of this work was to perform an homology modeling of the aaNAT, as well as searching for bioactive molecules which can complex with this target in order to act as inhibitors. Using the Swiss Model Workspace (https://swissmodel.expasy.org/), protein modeling results showed dopamine N-acetyltransferase protein (PDB code 3V8I) as the best template, with 60.10% identity and and 72% of corverage. The protein model was validated with a QMEAN value of -0.23. A virtual screening apprach was used to find ligand compounds which can complex with aaNAT, using ZINC database of natural and synthetic compounds. Autodock Vina calculations revealed several ligands with high affinity energy with aaNAT which can be proposed as new inseticides against A. aegypti. The best protein-ligand complexes will be subjected to molecular docking calculations for describing ligand behavior inside the active pocket for 50 nanosseconds of calculation.

Funding: CAPES

”’

Network Creation and Comparison From MicroRNAs Extracted From Peripheral Blood Of Primigravidae Submitted Or Not To Psychosocial Intervention

Rayssa Maria de Melo Wanderley Feitosa, Helena Brentani, Ariane Machado Lima, Gisele Rodrigues Gouveia

Universidade de São Paulo

Abstract

Environmental disturbance during the initial phases of human development, especially the gestational period, brings consequences that can last the offspring lifetime. Among these environmental problems, the mother's exposition to different stressors is correlated to a significant increase in the offspring's risk of developing various adversities, including cognition problems, emotional reactivity, impaired sociability and psychiatric disorders. The here proposed study is part of a randomized double-blind psicossocial intervention for primigravidae in socioeconomic vulnerability. Notwithstanding, there is a lack of knowledge on the possible biological markers of the intervention. Taking this into account, the present study has the goal to associate an important epigenetic factor, the microRNA, which has its expression associated to the period mentioned and disturbances along it, with the intervention. To this end, a comparison between two interaction networks microRNAs-mRNA will be performed. The complex networks are going to be created using the differentially expressed microRNAs from two gestational times. One network will be created using the differentially expressed microRNAs between the baseline (T0) and 30 weeks of gestation (T1), extracted from placental exosomes from peripheral blood of pregnants submitted to psychosocial intervention (cases), and the other network from differentially expressed microRNAs from the same moments (T0 and T1), collected from mothers not submitted to the intervention (controls). After the microRNA differential expression results from the RT-qPCR, a target prediction will be executed, using only experimentally validated targets, and the two networks containing the miRNAs and its target genes are going to be created on Cytoscape. The databases for experimentally validated targets are going to be extracted from miRTarBase, TarBase and miRwalk2.0. Different measurements collected from each network individually, using global topological properties and graph entropy are going to be compared to obtain a unique global value that represents the graphical differences between the two networks created.

Funding: Fundação Maria Cecília Souto Vidigal, Grand Challenges Canada, FAPESP e CNPq

,

Semantic Similarity Integration for Gene Network Inference

Roger Verzola Peres de Lima, Fábio Fernandes da Rocha Vicente

Federal University of Technology - Paraná

Abstract

Genes are fundamental elements in the dynamic of biological systems. Finding out how genes interact with the dynamic of biological systems may foster not only a better understanding of living beings, but also the possible genetic manipulations of said living beings with a specific aim. Therefore, the inference of gene regulatory networks is of great importance. However, an issue in that field is the small amount of samples available when compared to the amount of variables, severely limiting the inference power of purely statistical methods. In this work a method that contours that difficulty by uniting both quantitative data and qualitative data is proposed. Our method combines two types of data: gene expression and gene ontology. The criterion function calculates the mean conditional entropy over the normalized gene expression data and the GFD-Net over the genes' ontology annotations. GFD-Net is a method that gives a numerical score to the functional dissimilarity of a gene network based on gene ontology. The proposed algorithm to use is the Sequential Forward Feature Selection (SFFS) due to its easy implementation and deterministic nature. Therefore, the proposed method is made of two parts: an algorithm that selects a suboptimal set of genes (called predictors) that may interact with the target gene; and a criterion function that said algorithm will use to determine which subset of genes the predictor set will be. Running said algorithm over every target gene enables us to form a gene regulatory network by creating an edge between each predictor set and their respective targets. The method aims to use two distinct forms of evaluation unifying, therefore, both semantic and quantitative measures.

Funding: CAPES, CNPq, Fundação Araucária

”

Generative Adversarial Neural Networks for a Multiomics Approach in the Mycobacterium Tuberculosis Complex Analysis

Salvador Sánchez Vinces, Ana Marcia de Sá Guimarães, Ronaldo Fumio Hashimoto

University of São Paulo

Abstract

This work presents the application of generative Deep Neural Network methods to Mycobacterium tuberculosis (MTB) gene expression data (with preliminary results). These methods allow generating data with the same distribution as the original samples, as well as facilitating selective data generation of subgroups of original samples, and allowing some degree of manipulation to generate states of gene expression profiles. With such a variety of data, it is possible to establish further processing that facilitates analysis of genetic information (genome and transcriptome). The aim is to deepen the development of this new area of application of generative deep learning methods, studying characteristics and required preprocessing of the input biological data and optimizing the structures of neural networks for searching biologically plausible and integrated results at different levels of genetic information, and thus obtain data of interest in the amount required to make robust inferences using different models (e.g., identification and comparison of phenotypes by co-expression). For implementation and testing of the proposed model, we find convenient to work with MTB as reference microorganism within the MTB complex, for the relatively greater amount of information available, for example they have small but complex genome (approx. 4000 genes) and because their expression mechanisms are relatively well understood (approx. 40% of their genome has been characterized). The generated data (gene expression) were evaluated using qualitative distribution metrics such as histograms and t-SNE, and the effect on gene co-expression models. For both histograms and t-SNE, the generative model achieves a very similar distribution of values compared to original samples. Co-expression analysis shows a positive increase in the number of genes and modules inferred from the generated data when compared to the ones obtained from original data, such as modules neglected by the latter.

Funding: CAPES

”’

An integrated computational pipeline for inferring microbe-host interactions

Tahila Andrighetti, Leila Gul, Tamas Korcsmaros, Padhmanand Sudhakar

Earlham Institute

Abstract

Microbiota-host interactions are inherent in the evolution of most organisms with both positive and negative impacts. Hence, investigating host-microbiome interactions is crucial for understanding ecosystem dynamics, as well as the metabolism and physiology of diverse organisms. One way to evaluate such interactions is to study how organisms such as bacteria interact with their hosts at a molecular level. By detecting interspecies protein-protein interactions, it is possible to infer the host molecular mechanisms which are modulated by the bacterial proteins. However, detecting such interactions by experimental techniques remains challenging from a time and cost perspective. A more viable alternative to studying microbiome-host interactions is to use computational tools to predict them. In this work, we developed a pipeline by which it is possible to predict microbiome-host interactions and evaluate which molecular mechanisms in the host are potentially modulated by microbial proteins. With this end in sight, our pipeline integrates multi-omics data, such as metaproteomics which provides information about the composition of the proteins, microbe-host protein-protein interaction prediction along with host multilayer molecular networks. As a use case, we used a metaproteomics dataset which contains data (metaproteomics, metagenomics, host transcriptomics) from patients diagnosed with Crohn's disease (CD) and those who are healthy. By selecting differentially expressed proteins between the two conditions, we first predicted which bacterial proteins interact with human receptor proteins by using domain-domain and domain-motif interaction information from public databases. Then, we selected autophagy genes as potential target nodes, as autophagy is one of the known dysregulated cellular processes in CD. The next step consisted of compiling a signaling network which starts from bacterial protein-host receptor interactions, and ultimately reaching the selected host target genes through protein-protein and transcriptional regulatory interactions. From the obtained network, it was possible to identify putative molecular mechanisms by which bacterial proteins can modulate autophagy in the context of Crohn's disease.

Funding:

,

Fcoex: an R package for detecting co-expression modules in single-cell RNA-Seq data

Tiago Lubiana, HELDER T I NAKAYA

Institute of Mathematics and Statistics, University of São Paulo

Abstract

The boom of single-cell transcriptomics was followed by a growth in methods for the analysis of single-cell data. Currently, standard pipelines (such as Seurat and Biocondutor's OSCA) do not include co-expression network building and modules detection methods. Modern systems biology uses co-expression networks both for exploratory data analysis and gene regulatory network inference. Current methods for building these networks, such as WGCNA, were developed for bulk RNA-Seq and do not perform as well in single-cell data. In the present work, we show how a feature selection algorithm, the Fast Correlation-Based Filter (FCBF), can be used to detect co-expression modules in single-cell data via an R package called fcoex. The package is awaiting reviews for the Bioconductor repository and is available at https://github.com/csbl-usp/fcoex. We applied it to single-cell data from human, mice, and zebrafish, detecting o co-expression modules with known biological partners and putative associations. The presence of anticorrelated genes in the same modules allowed the detection, in the zebrafish dataset, of a module containing both a ligand (apela) and its receptors (aplnra/aplnrb), yielding insights into the biology of vertebrate development. Also, fcoex enables module-based reclustering of the datasets for multilevel labeling of cells, uncovering new populations, and avoiding trade-offs of the current label-determination methods. In parallel, we detected new candidates for subpopulations of zebrafish embryo cells and human blood monocytes, demonstrating the usefulness of our tools for exploratory data analysis of single cells.

Funding: This work was supported by the grant 2018/10257-2, São Paulo Research Foundation (FAPESP)

”

mirtronDB: a mirtron knowledge base

Bruno Henrique Ribeiro da Fonseca, Douglas Silva Domingues, Alexandre R Paschoal

UTFPR

,,,,,

Genomic analysis unveils important aspects of population structure, virulence, and antimicrobial resistance in Klebsiella aerogenes

Hemanoel Passarelli Araujo, Jussara Kasuko Palmeiro, Kanhu Charan Moharana, Francisnei Pedrosa da Silva, Libera Maria Dalla Costa, Thiago Venancio

,,,,,,,,

Prediction of new vaccine targets in the core genome of Corynebacterium pseudotuberculosis through omics approaches and reverse vaccinology

Carlos Leonardo Araújo, Jorianne Thyeska Castro Alves, Wylerson Nogueira, LINO CESAR DE SOUSA PEREIRA, anne cybelle pinto gomide, Rommel Thiago Jucá Ramos, Vasco A de C Azevedo, Artur Silva, adriana ribeiro carneiro folador

Universidade Federal do Pará

,,,,,,,,,,

Reverse vaccinology and subtractive genomics reveal new therapeutic targets against Mycoplasma pneumoniae: a causative agent of pneumonia

Thaís Cristina Vilela Rodrigues, Arun Kumar Jaiswal, Alissa de Sarom, Letícia de Castro Oliveira, Carlo Jose Freire Oliveira, Preetam Ghosh, sandeep tiwari, Fábio Malcher Miranda, Leandro de Jesus Benevides, Vasco A de C Azevedo, Siomar de Castro Soares

Universidade Federal do Triângulo Mineiro

””

nAPOLI: a graph-based strategy to detect and visualize conserved protein-ligand interactions in large-scale

Alexandre Victor Fassio, Lucianna Helene Silva dos Santos, Sabrina de A. Silveira, Rafaela Ferreira, Raquel Melo Minardi

Universidade Federal de Viçosa, UFV

””

Assessment of complementarity of WGCNA and NERI results for identification of modules associated to schizophrenia spectrum disorders

Arthur Sant'Anna Feltrin, Ana Carolina Tahira, Sérgio Nery Simões, Helena Brentani, David Correa Martins Jr

Universidade Federal do ABC

”’

The pH Signaling Transcription Factor PAC-3 Regulates Metabolic and Developmental Processes in Pathogenic Fungi

Maíra Pompeu Martins, Nilce M. Martinez-Rossi, Pablo R. Sanches, Antonio Rossi

”

N3O: A NEAT expansion for improving classification and feature selection applied to microarray data

Bruno Iochins Grisci, Marcio Dorn, Mario Inostroza-Ponta

Universidade Federal do Rio Grande do Sul (UFRGS)

”’

Uncovering the mouse olfactory long non-coding transcriptome

Antônio Pedro de Castello Branco da Rocha Camargo, Thiago Seike Nakahara, Marcelo Falsarella Carazzolle, Fabio Papes

Universidade Estadual de Campinas

,,,,,,,,,,

DETECÇÃO E VISUALIZAÇÃO DE SUBESTRUTURAS COMUNS NA INTERFACE PROTEÍNA-LIGANTE EM NÍVEL ATÔMICO ATRAVÉS DE MINERAÇÃO DE SUBGRAFOS FREQUENTES

Vagner Soares Ribeiro, Charles A. Santana, Alexandre Victor Fassio, Adriana M.Patarroyo-Vargas, Maria G. A. Oliveira, Valdete M. Gonçalves-Almeida, Sandro Carvalho Izidoro, Raquel Melo Minardi, Sabrina de A. Silveira, Pedro M Martins, Samuel da S. Guimarães, Sócrates Soares Araújo Júnior

Universidade Federal de Viçosa, UFV

Index of Authors