Proceedings X-Meeting 2017

Editor: AB^3C

October 2017

Conference Program

Highlight Track 165

1 | Organizing Committee

AB3C President : Alan M Durham (USP)

AB3C Vice President : Ney Lemke (UNESP)

AB3C Secretaries :

- Marcelo Brandão (Unicamp)
- Fabrício Martins Lopes (UTFPR)

AB3C Financial Department :

- Priscila Grynberg (Embrapa)
- Nicole Scherer (Fiocruz)

Poster Session Organizers :

- Robson Francisco de Souza (USP)
- Nicole Scherer (INCA)

Scientific Comitttee :

- André Fujita (USP)
- André Yoshiaki Kashiwabara (UTFPR)
- Arthur Griüuber (USP)
- Guilherme Targino Valente (Unesp)
- Katlin Brauer Massirer (UNICAMP)
- Márcio Dom (UFRGS)
- Maria Berenice Reynaud Steffens (UFPR)
- Robson Francisco de Souza (USP)

Highlight Track Organizer :

- Priscila Grynberg (Embrapa)
- Nicole Scherer (Fiocruz)

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 São Pedro on October 4^{th}-6^{th} .

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

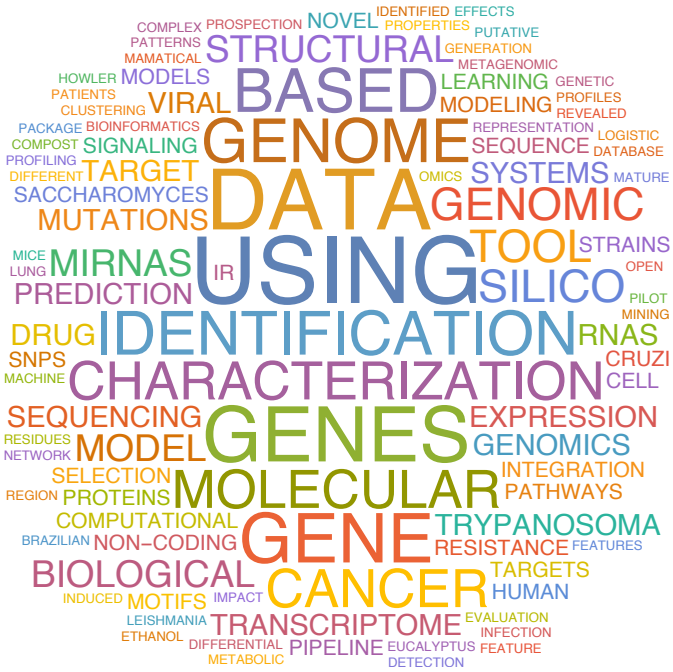

Figure 2.1: Word Cloud for the words used on the Conference Papers Titles

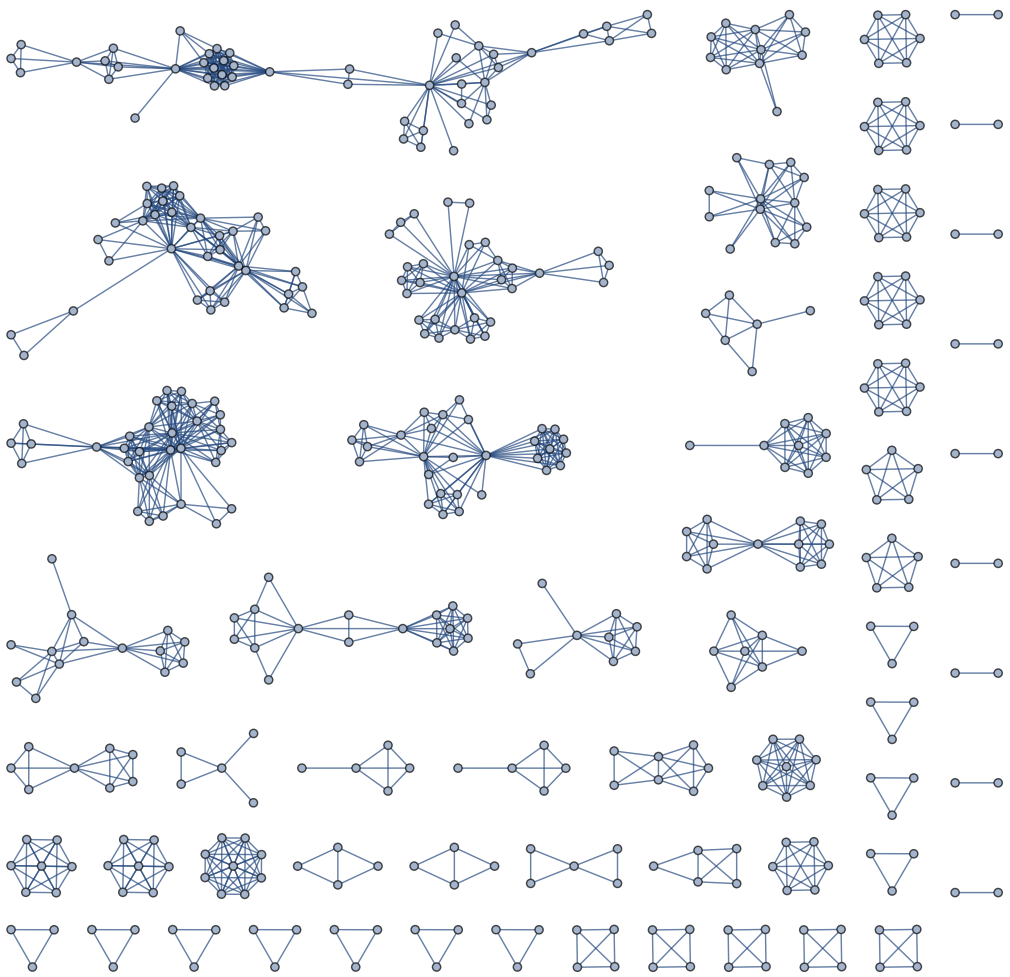

Figure 2.2: Graph representing the network of collaborations of the X-meeting 2016. An interactive version of this graph can be seen at https://neylemke.github.io/assets/grafo.html

3 | Genes and Genomics

Preliminary association of putative GBS-based SNPs with brown rust resistance phenotypes in a sugarcane map population using GWAS and machine learning methods

Alexandre Hild Aono[1], James Shiniti Nagai[1], Estela Araujo Costa[1], Hugo Rody Vianna Silva[1], Fernanda Raquel Camilo Dos Santos[2], Luciana Rossini Pinto[2], Anete Pereira de Souza[3], Reginaldo Massanobu Kuroshu[1]

1 UNIFESP
2 INSTITUTO AGRONÔMICO DE CAMPINAS
3 UNICAMP

Abstract

Brazil is the world's largest sugarcane producer and its production is an important source of income, arising mostly from sugar and ethanol. Currently, an approach for generating population data for this complex polyploidy species is genotyping-by-sequencing (GBS), which allows the detection of genomic variants without a reference genome. Several diseases limit sugarcane yield and the brown rust is considered one of the most important fungal diseases of sugarcane. Herein, putative GBS-based SNPs in a sugarcane map population were associated with brown-rust phenotype differences, being genetically linked to its causative factor. A set of 182 full-sibs derived from a sugarcane commercial cross between IACSP96-3018 and IACSP-3046 were generated and genotyped by GBS. Performing a comparative alignment (BWA version 0.7.15) of 831 million reads against methyl-filtered (MF) genome, we selected the correspondences with MF Coding-DNA sequences (CDSs). The data was pre-processed using Genome Analysis Toolkit (GATK) pipeline and the variants were called using HaplotypeCaller, implemented in GATK v3.7, and SAMtools v 1.4. After a series of stringent filters, we obtained a total of 8,345 SNPs and 290 indels, identified by both callers. In order to detect putative markers associated with brown rust resistance, a Genome Wide Association Study (GWAS) and machine learning approaches for feature selection were performed in the identified variants. Using an evaluation of the severity of brown rust, we classified the cultivars in two categories: most resistant and most susceptible. As a first step in the GWAS, the principal component analysis (PCA) was performed followed by the calculation of squared Euclidean distances between isolates and hierarchical clustering with complete linkage. The correction of population structure was based on the Discriminant Analysis of Principal Component (DAPC) implemented in the R adegenet package 2.0.0. The method used for association was the multivariate DAPC-based approach. In addition, three machine learning methods of feature selection were applied to the same data, improving the results by focusing on predictivity. The mutual information (MI) was obtained to measure the dependency between the loci and the phenotype. A model of Logistic Regression (LR) was built to select the loci related to the coefficients with high levels of influence under the model and Support Vector Machine (SVM) was used as a nonlinear classifier, where the attributes with most importance under the model were retained. Using the results of the four different approaches, a total of 241 potential variants were identified in 195 different MF scaffolds. All the consensus

Identification and analysis of thermophilic protein sequences in compost sequencing data

Amanda Rodrigues[1], Aline Maria da Silva[1], João Carlos Setubal[1]

1 USP

Abstract

Typical aerobic composting is a self-heating process in which microbial metabolism drives the temperature above 50?^{o}C, followed by sustained high temperatures between 60-80?^{o}C, and then followed by gradual cooling of the compost pile. Composting is widely considered as a promising source of novel thermostable enzymes, particularly those related to biomass degradation. Protein thermostability is an important aspect of protein biochemical and biotechnological research. Commonly, for chemical reactions, high temperature could increase reaction activity and decrease reaction time. Various mechanisms of thermostability were discussed in the literature, and many authors pointed to changes in amino acid composition as one of clearest manifestations of thermal adaptation. Indeed, it is well-known that enhanced thermostability is reflected in specific trends in amino acid composition. In this sense, it is important to develop a validated method that can predict the stability of a given protein from its primary sequence. In this work, machine learning techniques combined with amino acid composition and dipeptide composition were used to discriminate between thermophilic and non-thermophilic proteins from compost sequencing data. In order to obtain a reliable dataset, three different techniques (artificial neural network, random forest and SVM) were combined to discriminate the proteins sequences. The combination of these methodologies has achieved a specificity of 97% for the test dataset composed of thermophilic and non-thermophilic sequences with experimental evidence extracted from Uniprot. From the sequences classified as thermophilic, sequences related to the degradation of biomass will be selected for phylogenetic and completeness analyzes. Interesting sequences will be synthesized and heterologously cloned.

Funding: CAPES, Fapesp

CNV calling and its characterization in the Brazilian population

Ana Claudia Martins Ciconelle[1], Júlia Maria Pavan Soler[1]

1 INSTITUTO DE MATEMÁTICA E ESTATÍSTICA - IME/USP

Abstract

A copy number variation (CNV) occurs when the number of copies of a particular region of the DNA differs from two in autosomes or one/two in allosomes and has an important role in the genetic variability in humans. The effects of CNVs to human diseases are not yet known, although several diseases have been associated to this kind of polymorphism, such as uric acid and nervous system disorders. Motivated by the unknown influence of CNVs on anthropometric measurements and cardiovascular phenotypes and in collaboration with the Laboratory of Genetics and Molecular Cardiology at the Heart Institute/InCor-FMUSP, the primary aim in this project is to estimate the CNV from SNP array platforms and understand its transmission rate in family data and its association with complex phenotypes. This project also aims to understand the CNV distribution in Brazilian population and between family members. A pipeline was proposed for CNV calling from SNP Array data by reviewing softwares and packages that are available in the literature. Using the database from the Baependi Heart project, we analyzed the genotype and phenotype data from 80 families to identify the CNVs and to understand their association with height. From the genetic data, the CNVs were estimated from a combination of statistical techniques and algorithms including quantile normalization, classification methods for genotyping of SNPs and hidden Markov chains. After an exploratory data analysis, polygenic linear mixed models were used to estimate the association of CNVs with the chosen phenotypes. Our results suggest that the Brazilian population have a similar number of CNV per person (around 55 CNVs) in comparison with other populations. However, from the total of 64.107 CNVs identified, 147 CNVs are common among our samples, but rare in worldwide populations, one example being a CNV in the NEGR1 gene, which is present in 89% of our sample. Based on family data, the intraclass correlation coefficient for CNVs was estimated between 30% to 60% showing a high similarity on values from the same family. Association analysis between CNVs and different phenotypes are being performed.

Funding: CNPq, CAPES

Copy number variations of genomic and transcriptomic motifs of Mucin and MASP superfamilies in different Trypanosoma cruzi strains

Anderson Coqueiro Dos Santos[1], Gabriela Flavia Rodrigues Luiz[1], Najib M. El-sayed[2], Santuza Maria Ribeiro Teixeira[1], João Luís Reis Cunha[1], Daniella Bartholomeu[1]

1 INSTITUTE OF BIOLOGICAL SCIENCES, UFMG
2 DEPARTMENT OF PARASITE GENOMICS, INSTITUTE FOR GENOMIC RESEARCH

Abstract

Trypanosoma cruzi is the causative agent of Chagas disease, an illness that afflicts about 7 million people worldwide. Due to its extensive genetic variability, T. cruzi taxa is divided into six discrete typing units (DTUs), named TcI to TcVI. The first T. cruzi genome was sequenced in 2005, allowing the identification of hundreds of genes encoding polymorphic surface proteins from trans-sialidase (TcS), MASP and mucin (TcMUC) superfamilies. These genes are enrolled in cellular adhesion and invasion, and immune evasion processes, highlighting their important role in host parasite interactions. The high number of copies and variability of these gene families hinders the assignment of reads to specific genes. Members of these families share short motifs whose occurrence and abundance can be used to estimate the variability of these gene families among T. cruzi strains. In the present work, we aim to compare the motif amplifications in the multigene families MASP, TcMUC and TcS, derived from representatives of T.cruzi TcI, TcII and TcVI DTUs, and compare their copy number with gene expression levels. We used genomic reads from two clones derived from TcVI CL strain (CL Brener, CL-14), Y strain (TcII) and 3 representatives from DTU TcI; as well as transcriptomic reads from the same clones/strains in the amastigote, epimastigote and trypomastigote stages. These genomic and transcriptome reads were mapped in CL Brener genome, and only reads that mapped in the TcS, TcMUC and MASP genes were recovered. Kmers of 30 nucleotides were generated from these reads and their coverage was estimated. To remove redundancy, similar kmers were clustered and the number of motifs, for each library, was normalized to allow comparisons among them. For TcMUC superfamily, the analysis showed a greater number of motifs in the virulent CL Brener clone, compared to a reduced number of motifs in the non-virulent CL-14 clone. We did not detect large differences within the three clones derived from Y strain and little variance in TcI genomic motifs. When we compared the different developmental stages, a greater concordance was found between the amastigote 60 and 96 hrs from CL-14, whereas in CL Brener the pattern observed in amastigote 96 hrs and trypomastigote stage was quite similar. Similar results were observed for MASPs. We are currently performing this analysis with TcS family and comparing the motifs with higher counts in genomic and transcriptomic analysis to try to correlate copy number of the identified motifs and their expression levels.

Funding: CAPES, CNPq and FAPEMIG

The Pan-Genome of Treponema pallidum Reveals Differences in Genome Plasticity between the subspecies

Arun Kumar Jaiswal[1], Sandeep Tiwari[2], Syed Babar Jamal Bacha[2], Vasco A de C Azevedo[3], Siomar de Castro Soares[4]

1 INSTITUTE OF BIOLOGICAL SCIENCE, UFMG, BELO HORIZONTE; DEPARTMENT OF IMMUNOLOGY, MICROBIOLOGY AND PARASITOLOGY, INSTITUTE OF BIOLOGICAL SCIENCES AND NATURAL SCIENCES, UFTM
2 INSTITUTE OF BIOLOGICAL SCIENCE, UFMG, BELO HORIZONTE
3 UFMG
4 DEPARTMENT OF IMMUNOLOGY, MICROBIOLOGY AND PARASITOLOGY, INSTITUTE OF BIOLOGICAL SCIENCES AND NATURAL SCIENCES, UFTM

Abstract

Spirochetal organisms of the Treponema species are responsible for causing Treponematoses. Pathogenic treponemes cause multi-stage infections like endemic syphilis, venereal syphilis, yaws and pinta. Out of these four lethal diseases, venereal syphilis is transmitted by sexual contact; the other three diseases are transmitted by close personal contact. Treponema pallidum subspecies pallidum is Gram-negative, motile, spirochete pathogen that cause syphilis in human. Syphilis is a multistage infectious disease that can be communicated by sexual contact. The current worldwide prevalence of syphilis emphasizes the need for continued preventive measures and strategies. Unfortunately, effective measures are limited. The genome sequence of all 49 T. pallidum strains available from NCBI, isolated from different hosts and countries, were comparatively analysed using pan-genomic strategy. Phylogenomic, pan-genome, core genome and singleton analyses disclosed the close connection among all strains of the pathogen Treponema pallidum. The pan-genomic analysis showed that all the strains are highly clonal. Furthermore, the genome plasticity analysis among the subspecies T pallidum subsp pallidum, T. pallidum subsp endemicum and T. pallidum subsp pertenue revealed differences in the pathogenicity island (PAIs) and genomic island (GIs) repertoire. We found 4 pathogenicity island (PAIs) and 8 genomic island (GIs) in subsp pallidum, whereas subsp endemicum has 3 PAIs and 7 GIs and subsp pertenue harbour3 PAIs and 8 GIs. The differences observed in genome plasticity among sub species can be useful for further characterization oftheir epidemic behaviour.

Funding: TWAS.CNPq and CAPES

Homology detection using multilayer maximum clustering coefficient

Caio Rafael do Nascimento Santiago[1], Luciano Antonio Digiampietri[1]

1 USP

Abstract

Sequence clustering is an important tool for helping the understanding of homology relations in protein sets, by grouping them into related families. The identification of these families is not a trivial task, and there are many studies dedicated to solving it, the majority of them are graph-based ones. In the graph-model, nodes represent sequences and edges represent homology relations, in general, defined considering metrics obtained from local alignments.

Some studies use the concept of transitivity to explain the homology, i.e., if two proteins are homologous and a third is homologous to one of the two, then the three are considered a family of homologous proteins. The main concern about this approach is the establishment of how much this proposition could be extended based only on transitivity.

This work presents a graph-based clustering method that maximizes the clustering coefficient based entirely on the transitivity of the homology relationship. It creates a undirected graph considering the e-value metric (or any other alignment metric chosen by the user) and produces a multilayer list of gene families according to thresholds progressively more restrictive. This allows to the user to work with genes families composed of genes with greater distances (first layers) and more restricted families (with more similar genes) in the last layers.

Some advantages of this approach are: it preserves the graph constructed considering the local alignments and, therefore, it is easy to understand the process that generated a certain family; it is possible to analyze the topology of the graph, in order to, for example, find multi-domain proteins or find the proteins that phylogenetically separate two families.

This approach was tested in two phylogenetically closely related sets of genomes, the first contains 69 strains from Xanthomonadaceae family and the second contains 55 Streptococcus pyogenes strains. The results from our approach were compared with the TribeMCL results. In the case studies, our solution identified a bigger core genome, considering the number of homologous families (4% bigger than the one identified by TribeMCL), and, when ignoring the paralogs genes, our approach identified a core with 42% more homologous families than the one identified by TribeMCL. When analyzing the gene annotation/products in the homologous families, our solution was 6% better in grouping genes with the same annotations in the same family when compared with the families produced by TribeMCL and using the annotations provided by Patric. Finally, the biggest families produced by our approach are smaller than the ones produced by TribeMCL (from 9% to 15% smaller), being able to not group together genes that are not very similar and have different annotations.

Funding: Capes

New approach for genomic comparison of invasive and non-invasive strains of Streptococcus pyogenes

Suzane de Andrade Barboza[1], Caio Rafael do Nascimento Santiago[1], Luciano Antonio Digiampietri[1]

1 USP

Abstract

Streptococcus pyogenes or Group A streptococcal (GAS) is a uniquely human Gram-positive pathogen related to a wide range of invasive and non-invasive diseases, having the fourth highest mortality rate among bacterial pathogens. The diversity of clinical outcomes of these infections can be explained by the acquisition of exogenous genetic material, mostly composed of virulence factors such as adhesins or phage toxins. One of the main virulence factors is M protein, which hypervariable region is used for GAS classification. Molecular epidemiology studies showed a genotype M/pathogenicity relation, which is being intensively investigated by genomic comparisons. However, little is known about different invasive levels observed within strains sharing the same genotype. This lack of information occurs due to two main reasons: (1) recent studies limit their comparisons by genotype or pathology analysis, disregarding non-invasive strains, and (2) software limitations concerning closely related genomes comparisons, which include difficulties in performing global alignments considering large genome rearrangements and the identification of strains' exclusive genes. In order to overcome these difficulties, two main strategies have been used in the comparison of 55 GAS genomes (28 genomes from invasive strains, 25 from non-invasive strains and 2 from isolates with unknown invasive profile): phylogenetic and gene network analysis, based on the identification of homologous genes. After performing local alignments with all genes of all genomes, homology relations were defined considering seven defined parameters: minimum identity percentage, minimum alignment percentage, minimum alignment length, maximum number of mismatched positions, maximum number of gap positions, maximum e-value, and minimum bit-score. The resulting graph is a gene network representation, where each homologous gene group is a cluster composed of nodes representing the genes of the genomes. Each genome is represented by a color, which allow us to identify gene sets exclusively found on invasive strains or strains related to a certain disease. A distance matrix of the genomes has then been calculated based on presence or absence of genes for each group of genes created previously, and a cladogram (representing the phylogenetic relationships) of all genomes was constructed, grouping strains with similar gene composition. Relating this information with the disease/virulence profile, we aim to better understand the relation between GAS genotypes and gene acquisition. These graphical representations will accelerate the identification of the virulence factors that could explain certain isolates' invasiveness and alternative genes for the production of an anti-streptococcal vaccine.

Funding: Capes

Development of an integrated genetic map for a full-sib progeny from crossing between Eucalyptus grandis and Eucalyptus urophylla

Cristiane Hayumi Taniguti[1], Izabel Christina Gava de Souza[2], Shinitiro Oda[2], Leandro de Siqueira[2], Rodrigo Neves Graça[3], Thiago Romanos Benatti[2], José Luiz Stape[2], Antonio Augusto Franco Garcia[1]

1 USP
2 SUZANO PAPEL E CELULOSE
3 FUTURAGENE

Abstract

Brazil is among the largest eucalyptus producers in the world. The culture has great importance in Brazilian economy, attending a variety of markers, as the cellulose market. The commercial importance of this crop requires constant improvement of the cultivars. The increasing improvement and availability of the genotyping and phenotyping high-throughput platforms consolidate promising proposals to accelerate eucalyptus breeding programs. The data generated by such platforms are used for genetic understanding of quantitative traits, which are the majority of the commercially-targeted characteristics. Linkage maps are fundamental tools for analyzing these characters. However, it is necessary the development of new strategies for the construction of genetic maps adapted to high-throughput data. These strategies also requires considering particular genetic aspects of eucalyptus species, as their outcrossing breeding system, which makes available only F1 mapping populations. These populations can have greater number of alleles per locus and unknown linkage phases compared to inbred based populations. Furthermore, molecular characteristics obtained by eucalyptus genome sequencing can be useful to linkage map building process. The aim in the present work was to construct an integrated genetic map in a full-sib progeny with 200 individuals, derived from the cross between Eucalyptus grandis and Eucalyptus urophylla. For markers identification, it was performed a complete genome re-sequencing (WGS) of the parents and genotyping-by-sequencing (GBS) of the progeny. The mapping methodology was adapted to the data set, which presents a large amount of SNP markers, only containing diallelic information and with variable genotyping error probability. For this, two strategies were proposed: i) use of genome reference information as aid for map construction; ii) adapting the genotyping error probability parameter in the approach implemented in the software OneMap. The map presented a total size of 1471.91 cM and 1512 markers, with a mean distance between them of 1.85 cM. The markers formed 11 linkage groups, which corresponded to chromosomes of the reference genome. On average, 96.8 % of chromosomes were covered. The obtained map showed recombination rate pattern similar to other maps constructed for eucalyptus. Also, 61 markers located in other scaffolds in the reference genome were grouped with the 11 groups and they may serve to elucidate the assembly of these. Using the proposed strategies, a suitable integrated map was obtained for the present experiment.

Funding: CNPq, Suzano Papel e Celulose

Gene expression biclustering with FIrefly Algorithm

Denilson Oliveira Melo[1], Paulo Eduardo Ambrósio[1]

1 UESC

Abstract

Gene expression profiling is the measurement and study of expression levels of various genes at once, with the intent to discover and understand gene function. Gene expression data is usually presented in an expression matrix, which can be analyzed with computational and statistical methods. One way to analyze this expression data is through clustering techniques, which aim to group genes of similar expression tendencies together, suggesting that these genes are subject to a common pattern of regulation. Clustering techniques however are limited, they are only able to create groups considering the entire set of conditions, but genes are not necessarily related to every condition, and they also doesn't allow coupling, thus ignoring the possibility that an individual gene may be related to multiple groups in different subsets of conditions. This limitation can be surpassed with biclustering techniques, which aim to group both genes and conditions together. With biclustering groups are formed considering each subset of condition, thus allowing more grouping possibilities that better represent the relationships between genes. Biclustering however is a NP-hard problem. Problems of this class have no exact solution and are usually solved using heuristics. Several metaheuristics are already being used to find biclusters and in this preliminary study we propose to explore and utilize one of nature-inspired metaheuristic proposed by Xin-She Yang - the Firefly Algorithm (FA) - a swarm based approach inspired from the behaviour of fireflies revolving around their blinking patterns. This metaheuristic considers the attraction between fireflies, where given two individuals, the less bright one will be attracted to the brightest. With little effort one could see that with these parameters groups of fireflies would be formed naturally, and that's the basic idea. In this preliminary study we aim to employ the grouping characteristics of the algorithm to find biclusters of gene expression data.

Funding: Fundação de Amparo à Pesquisa do Estado da Bahia - FAPESB

MARVEL: A pipeline for recovery and analysis of viral genomes from metagenomic shotgun sequencing data

Deyvid Amgarten[1], Aline Maria da Silva[2], João Carlos Setubal[2]

1 USP - DEPARTAMENTO DE QUIMICA
2 USP

Abstract

The study of the viral diversity in environmental samples has become increasingly important due to the recognition of key roles played by these organisms in diverse ecosystems. Recent works provide evidence that viruses of bacteria (bacteriophages) are key players in biogeochemical cycles of large ecosystems, such as oceans and forests. Viruses may also be determinant in the flux of genes among microbial populations and in the plasticity of microbial communities, helping these communities to deal with environmental stresses. Knowing the genomes of viruses that are present in diverse environments can thus help the understanding of the microbial ecology and evolution of these environments. Here we describe the MARVEL pipeline for recovery and analysis of viral genomes from metagenome shotgun sequencing data. The main steps in this pipeline are: sequence quality control, metagenome assembly, similarity searches against wide and hallmark-protein databases of viruses, removal of false positives, and multisample contig binning. At the end, MARVEL generates an automatically curated set of contigs that correspond to draft and complete genomes of environmental viruses present in the analyzed sample. We have applied MARVEL to metagenomic datasets obtained in two environments (composting and a reservoir) of the Sao Paulo Zoo. We obtained 37 viral genomes from reservoir samples and 36 viral genomes from composting. Most of these genomes have low or no similarity with viral genomes in public databases. Therefore these results are a contribution for shedding light on the gigantic viral dark matter that exists in our planet. MARVEL can be applied to any shotgun metagenomic dataset for which Illumina reads are available.

Funding: Funding for this research is provided by FAPESP (2014/16450-8) and CAPES

Biovar equi versus ovis: What genetically differentiate them?

Doglas Parise[1], Mariana Teixeira Dornelles Parise[1], Marcus Vinicius Canário Viana[2], Elma Lima Leite[3], Anne Cybelle Pinto Gomide[2], Vasco Ariston de Carvalho Azevedo[1]

1 UFMG

Abstract

Corynebacterium pseudotuberculosis is a pathogenic bacteria that causes significant economic loses in global livestock. This organism is classified into biovars ovis and equi. Strains from biovar ovis are more clonal and infect small ruminants, such as goats and sheep, and eventually humans. Strains from biovar equi infect bigger animals such as bovines, buffaloes and horses causing different diseases. This study aims to extend a previous work in which genes that could differentiate the biovars equi and ovis were identified in six C. pseudotuberculosis genomes isolated in Mexico. All the 53 complete and non-redundant genomes available from NCBI were used. The Bacterial Pan Genome Analysis (BPGA) pipeline was used to estimate the pangenome and find specific genes of genome subgroups. Groups of orthologs were defined with a similarity cutoff of 70%, using USERACH tool. The results show a pangenome of 2,290 and a core genome of 1,277 genes. The core genome can be used to identify new targets for vaccine and diagnosis methods. The total number of biovars exclusive genes is 265, 65 of them from biovar ovis (15 with known function) and 200 from biovar equi (54 with known function). As expected, we identified genes known to differentiate the biovars, such as the nitrate reductase operon narGHIJ. In addition, the results corroborate the previous study with Mexican strains that pointed CRISP-Cas genes and a restriction endonuclease type III (Restriction Modification System) as exclusive from biovars equi and ovis, respectively. The remaining genes will be further studied.

Funding: Capes, CNPq e Fapemig.

Ploidy level analysis of functional SNPs from GBS data in a sugarcane map population

Estela Araujo Costa[1], Alexandre Hild Aono[1], Hugo Rody Vianna Silva[1], James Shiniti Nagai[1], Anete Pereira de Souza[2], Reginaldo Massanobu Kuroshu[1]

1 UNIFESP
2 UNICAMP

Abstract

Sugarcane is one of the most important energy source in Brazil. Because of the complexity in determining ploidy levels, studies with molecular markers have been limited by the lack of information so far. Genotyping-by-sequencing (GBS) made possible deep genetic analysis, making accessible studies with molecular markers using ploidy level from genomic data of complex polyploids. Herein, a pipeline for identifying SNPs from GBS data in poliploidy genomes was established and we used the SuperMASSA software to estimate the ploidy level of some functional SNPs. In total 831 million 100-bp single end reads were generated from GBS of 182 individuals from a sugarcane map population. After demultiplexing and barcode processing, BWA version 0.7.15 was used to align reads against sugarcane methyl-filtered (MF) genome sequence. From a total of 96% of reads aligned, we selected alignments found in 10,957 MF Coding-DNA sequences (CDSs). Variants were called using GATK and Samtools, resulting in 8,345 SNPs in the intersection of both callers. All consensus loci were compared to Sorghum bicolor, Oryza sativa and Zea mays CDSs genomes using BLASTn. An enrichment analysis was performed where 538 SNPs in 94 MF scaffolds, which correspond to some important GO categories involved in sugar transport and metabolism in the storage tissue. Those 94 scaffolds were used as input to KAAS server, resulting in 65 scaffolds with 352 SNPs representative in 64 KEGG orthology (KO) terms. During KEGG analyses, some important pathways were represented, such as "Glycolysis / Gluconeogenesis" and "Amino Sugar and Nucleotide Sugar Metabolism", with four and five SNPs respectively. In order to genotype the population, the SuperMASSA software was used to infer on the ploidy level in a range from 2 to 20, and a posterior probability was associated. In total, 510 putative SNPs were able to estimate the ploidy level using SuperMassa. The results showed a large number of ploidy level at 20 (35%) in the category GO:1901576 'organic substance biosynthetic process'. The most representative ploidy level was 20 (50%). These results suggest a high level of ploidy estimation. Possibly, because those called SNPs are in genes involved to important energetic process such as the sugar metabolism, they might have multiple copies in the sugarcane genome, what makes the estimative performed by the SuperMassa software biased towards high ploidy levels. By unveiling many putative functional SNPs, the pipeline we established brought positive results to help the understanding of the complex ploidy levels of sugarcane genome.

Funding: Capes

Lower proportion than expected for transversions and higher for transitions in synonymous SNPs evaluated in dbSNP

Fernanda Stussi[1], Tetsu Sakamoto[2], José Miguel Ortega[2]

1 UFMG
2 UFMG, LABORATÓRIO DE BIODADOS

Abstract

We have previously investigated the proportions of SNPs involving transversions (A/C, A/T, G/C, G/T) and transitions (A/G and C/T) for human entries in dbSNP. Here we reanalyzed human SNPs together with mouse, rat, pig and cow deposits. For these organisms, SNPs were classified after the region where the SNP is located: Intron, 5'-UTR, 3'-UTR, CDS-missense and CDS-synonymous. We noticed that the deposit for cow SNPs suffered somehow a bias that changed the relative proportions between these categories, suggesting that sampling is not random in the cow project. Therefore we set up to investigate the proportions of transitions and transversions for the other four organisms. The frequency of transitions varied around 33% for all categories but CDS-synonymous, where it was over 40%, and this measurement was attained for man, mouse, rat and pig. Accordingly, all four transversions were lower for CDS-synonymous. C/G Transversions were also lower for Intron and 3'-UTR and especially higher in CDS-missense and 5'-UTR. Thus, these studies support the occurrence of equal amounts in non-coding regions of genome for transitions and all transversions but C/G. And furthermore suggest that using SNP frequency in non-coding transcript regions in man, rat, mouse and pig would be safe for building mutation models based on SNP frequencies.

Funding: CAPES/Biocomp

Comparative genomics of six Pseudomonas phages isolated from composting

Fernando Pacheco Nobre Rossi[1], Deyvid Amgarten[1], João Carlos Setubal[2], Aline Maria da Silva[2]

1 USP - DEPARTAMENTO DE QUIMICA
2 USP

Abstract

Bacteriophages (or simply phages) are viruses that infect bacterial cells and are the most abundant and, potentially, the most diverse biological entities on Earth. More than 1023 infections by phages are expected to occur every second. The dynamics of phage-host populations presents complex relationships and is thought to contribute to bacterial abundance and diversity as well as to environment homeostasis. In the billion years that phages co-existed with their bacterial hosts, phages have evolved highly diverse proteins that either inhibit or adapt bacterial metabolic processes to their own benefit. Since their discovery, in the early 20th century, phages and their proteins have been exploited as valuable molecular biology and biotechnology tools. Phages have been also considered potential antibacterial agents, and their use to reduce or eliminate bacterial infections is known as phage therapy. Phages might be a treatment option for?antibiotic resistant?bacteria. Phages that are lytic and that are not capable of displaying lysogeny are preferred for phage therapy purposes. In a previous work from our group, composting samples from the Sao Paulo Zoo Park were screened for phages infecting Pseudomonas aeruginosa PA14. Six phages were isolated and had their genome sequenced. One of them (ZC01) was shown to be from Siphoviridae Yu-A like genus and the other two (ZC03 and ZC08) were similar to each other and shown to be novel Podoviridae phages. All three phages are lytic and have the ability to degrade P. aeruginosa PA14 biofilm, and as such they can be promising candidates for antimicrobial application. In the present work, we extend this prior study by analyzing the three remaining phages (ZC04, ZC06 and ZC07). All three were predicted to belong to the Podoviridae family. Phylogenetic trees were generated based on multiple alignment of the terL marker gene using MAFFT, followed by alignment curation with GUIDANCE2 and maximum likelihood computation using RAxML. This analysis shows that phages ZC03, ZC08, ZC04, ZC06 and ZC07 are phylogenetically close. ZC06 and ZC07 are closer to each other than to others and their genomes have 99% similarity. The genomes of ZC04 and ZC03 have 97% similarity. Differences in these five genomes include INDELS that resulted in truncation of at least one CDS in ZC04. Other genomic differences that might have functional implications will be discussed.

Funding: Funding for this research is provided by FAPESP and CAPES

Analysis of genomic islands of virulence and pathogenicity in Xanthomonas campestris

Juan Carlos Ariute[1], João Pacifico Bezerra Neto[1], Ana Maria Benko-iseppon[1], Flavia Figueira Aburjaile[1]

1 UFPE, CENTER OF BIOLOGICAL SCIENCES, GENETICS DEPT.

Abstract

Brazilian viticulture activity has remarkably increased in the northeastern region of the country over the last years. However, significant losses have occurred due to the local climate rough conditions that raise susceptibility of grape vine species such as Vitis vinifera to bacterial infections. In this context, Xanthomonas campestris pv. viticola, a Gram negative aerobic pathogenic bacteria, is associated with bacterial canker and many other phytopathologies, due to factors such as xanthomonadin and xantham gum production. Since it was first isolated from an Indian viticulture in the 70's, extensive studies have tried to elucidate how to control X. campestris pv. viticola infections in grape vine. Nevertheless, there is still a shortage of information regarding the molecular mechanisms and the acquisition of virulence genes in this organism. Therefore, we believe that identifying and exploring genomic content from Xanthomonas campestris strains would improve knowledge concerning its pathogenicity and help develop future approaches to control the disease. Pathogenicity islands (PAI) are described as a set of virulent genes which have been horizontally transferred among Eubacteria, providing the organism with the ability to cause pathologies. In this sense, we aimed to identify genes that are present in PAIs from X. campestris pv. viticola. For automatic annotation process, seven complete genome of X. campestris pv. campestris strains obtained from National Center for Biotechnology Information (NCBI) genome database were submitted on Rapid Annotation using Subsystem Technology (RAST). Afterwards, the annotated genomes were analyzed on Seed viewer and Artemis. Finally, Genomic Island Prediction Software (GIPSy) and Islandviewer4 were used to provide a better prediction view and analysis of PAIs. We show that almost 45% of genome can be annotated using subsystems and that 77 genes are directly related to virulence, defense mechanisms and disease activity. The proteins encoded by these 77 genes were divided in two basic subcategories: antibacterial peptides and resistance to toxic compounds and antibiotics related to copper homeostasis. In summary, we reveal that the genomes of all seven Xanthomonas strains are very similar to each other but do possess relevant differences in terms of genes encoding pathogenicity factors. Moreover, there are high expectations for finding a resembling pattern on X. campestris pv. viticola, making possible the development of genetic improvements of plants, without the use of chemical agents.

Funding: CAPES, FACEPE, CNPq

Comparative genomics of Xanthomonas spp. focusing on CAZymes associated with host-pathogen specificity

Gabriela Persinoti[1], Mario Tyago Murakami[1]

1 CTBE/CNPEM

Abstract

Xanthomonas is a genus of Gram-negative Gammaproteobacteria that cause infections in leaves and fruits of plant hosts. Around 400 plants may be infected by Xanthomonas species, among them, many are economically important ones, such as citrus, rice, tomato and banana. A high degree of host specificity is observed among Xanthomonas pathogenic species and pathovars. For instance, X. citri pv. citri exclusively infects citrus while, other pathovas such as X. citri pv. mangiferaeindicae infects mango and X. vesicatoria may infect tomato and pepper. The goal here was to perform a large scale comparative genomics analysis of Xanthomonads focusing on the relationship of CAZyme-genome content and host-bacterium specificity. For the comparative genome analysis, 51 complete genomes of Xanthomonads species and panthovars were used. To investigate the phylogenetic relationship of Xanthomonads, 699 single copy genes with members in 51 species were identified. Single-gene alignments longer than 100 residues after excluding low-scoring alignment sites were concatenated into a supermatrix. The resulting supermatrix is composed of 111,390 distinct alignment patterns. Maximum likelihood trees were estimated using either FastTree using WAG+CAT model and RAxML using a distinc model for each of the 582 partitions. Both trees presented strong support values (Bootstrap > 95%) and the same topology, which was assessed by Robinson-Foulds distance implemented in phangorn R package. Two distinct groups were clearly formed among Xanthomonas species. One group is composed by X. sacchari, X. albilineans, X. translucens, and X. hyacinthi, while another larger group is composed by X. campestris, X. arboricola, X. gardneri, X. hortorum, X. fragariae, X. cassavae, X. bromi, X. oryzae, X. vasicola, X. fuscans, X. citri, X. axonopodis, X. euvesicatoria and, X. perforans including several pathovars. Regarding the CAZyme-genome content, X. fuscans and X. hyacinthi present, respectively, the fewer and greater number of CAZymes identified. For instance, X. citri pv. citri, a mesophyllic pathogen, which infects the intercellular spaces of the mesophyll tissue causing citrus canker, presents 239 cazymes, being many organized as PUL (Polysaccharide Utilization Loci). In the other hand, Xylella fastidiosa, also a citrus plant pathogen, but a vascular pathogen, which infects the xylem elements of the vascular system, presents only 82 CAZymes, and none of them are organized as PULs. This suggest that the CAZyome may play a role in host-pathogen interactions in Xanthomonads.

Funding: FAPESP and CNPq.

GBKFinisher: A tool for GenBank files refinement

Gustavo Santos de Oliveira[1], Doglas Parise[1], Mariana Teixeira Dornelles Parise[1], Anne Cybelle Pinto Gomide[2], Vasco Ariston de Carvalho Azevedo[1]

1 UFMG

Abstract

After genome assembly and annotation, usually a lot of effort is taken to check for potential pseudogenes as well as to set annotation features ready to submission. Those processes are usually manually performed in Artemis and are time consuming. Different filters set in Artemis help in the identification of potential issues, as absence of stop codons in CDSs, incorrect start codons, duplication and overlap of features, etc. For potential pseudogenes, however, a manual inspection must be performed. The user must check for frameshifts corrections by observing for possible indels, the presence of premature stop codons or even gene fragments. The user, then frequently uses BLASTp as a guide for possible correction of those different issues. Based on the need for a more automatic process we introduce GBKFinisher. This tool was developed as an effort to help users to save time and more accurately fix annotation issues. GBKFinisher is based on three major packages: GBKChecker, GBKParser and GBKSolver. GBKChecker attempts to give the basic stats, like qualifiers count and nucleotide composition, as well as a diagnosis of possible issues, like the number of possible frameshifts, fragments and split ORFs that could be safely joined, without a stop codon in between. GBKParser was created to rewrite gbk files with qualifiers as specified by the user. GBKSolver attempts to automatically solve for detected issues by GBKChecker. GBKFinisher allows a very convenient way for users to filter out unwanted qualifiers, as well as to annotate locus_tag differently for CDSs, tRNAs, and rRNAs. It is also possible for users to manually correct potential issues detected in GBKChecker. A log file with all possible issues encountered by GBKChecker is maintained and can be accessed for detailed troubleshooting. Likewise, a gbk file is created with a color scheme that assists users in the curation process. The primary output of GBKFinisher is a user defined Genbank file, that can ultimately be used for guiding the curation process or even for sequin and GenBank submission.

Funding: Fapemig

Analysis of metagenomic data from howler monkeys feces

Italo Sudre Pereira[1], Raquel Riyuzo de Almeida Franco[2], Layla Martins[1], Julio Oliveira[3], João Carlos Setubal[2], Aline Maria da Silva[2]

1 USP
2 UNIFESP

Abstract

There is an increasing use of metagenomic approaches to characterize microbial communities in several environments regarding their structure, function and composition, in particular to access the vast amount of uncultured microorganisms, enabling the understanding of the biological functions that such organisms play in these environments. Here we describe results of a project aiming to sample and analyze the fecal microbiota of captive and non-captive howler monkeys in the São Paulo Zoo. A previous study from our group using 16S rRNA amplicon sequencing has demonstrated differences in the microbiota profile between captive and non-captive individuals. In this project, we performed shotgun sequencing of total DNA obtained from thirteen samples from captive and six from non-captive monkeys, the samples were collected in different seasons. The diet of those groups are different and we have well detailed information about captive diet. Preliminary results show differences between the two groups in terms of taxonomic groups as well as in the function profile. The taxonomic results shows that Bacteroides, Prevotella and Parabacteroides are the most abundant genera in non-captive monkeys and Prevotella, Bacteroides and Clostridium are the most abundant genera in captive monkeys. The taxonomic and functional profiles results suggests that the non-captive monkeys are more susceptible to season changes and captive ones have a more homogeneous microbiota over the year.

Funding: FAPESP, CAPES, CNPq

Identification of genes under positive selection in Corynebacterium pseudotuberculosis

Marcus Vinicius Canário Viana[1], Henrique Figueiredo[2], Felipe Luiz Pereira[2], Fernanda Alves Dorella[2], Anne Cybelle Pinto Gomide[1], Alice Rebecca Wattam[3], Vasco A de C Azevedo[1]

1 UFMG
2 NATIONAL REFERENCE LABORATORY FOR AQUATIC ANIMAL DISEASES OF MINISTRY OF FISHERIES AND AQUACULTURE, UFMG
3 BIOCOMPLEXITY INSTITUTE OF VIRGINIA TECH, VIRGINIA TECH, BLACKSBURG, VIRGINIA, UNITED STATES OF AMERICA

Abstract

Corynebacterium pseudotuberculosis is a Gram-positive, intracellular pathogen, close related to the diphtheria etiological agents C. diphteria and C. ulcerans. The biovar Ovis infects mainly small ruminants causing Caseous Lymphadenitis, while biovar Equi infects larger animals, causing different diseases. The species virulence mechanisms and biovar differentiation are not fully understood, and drugs and vaccines are not effective for all hosts. The goal of this work is to identify genes under positive selection in C. pseudotuberculosis to better understand its evolution and collaborate with the development of control measures. Due to computational costs, 29 strains (16 Ovis and 13 Equi) from different hosts and countries were selected for a genome scale analysis using the POTION pipeline. We used FastOrtho for ortholog assignment, a cutoff of 50% for sequence identity within ortholog groups, PRANK for sequence alignment, and dnaml for phylogeny reconstruction. For the positive selection analysis, the pairs of null/positive selection site models M1a/M2a and M7/M8 were compared. Eight genes were identified: uncharacterized sigma 70, adhesin, manganese ABC transporter, fatty acid synthase, lambda repressor-like, tyrosine-tRNA ligase, and two uncharacterized transmembrane and secreted proteins. In addition, 14 other genes had evidence of recombination, including sialidase, ferrochelatase, adhesin and sortase. These proteins are related to metabolism and host colonization processes as adhesion, metal uptake, protein synthesis and gene regulation. The adaptations provided by the identified mutations will be investigated and a preliminary analysis of protein sequences shows correlation between biovar and specific amino acids.

Funding: CAPES, CNPq, FAPEMIG, UFMG

Identification of motifs in the promoter region of genes related to the ABA-dependent pathway in sugarcane

Mauro de Medeiros Oliveira[1], Alan Durham[1], Glaucia Souza Mendes[1]

1 USP

Abstract

In general, the promoter region consists of different regulatory elements, such as Transcription Factor Binding Sites (TFBS), which are responsible for the activation of gene transcription. TFBSs can be characterized using different experimental processes such as Chip-Seq. However, these experimental present low reproducibility for non-model organisms. For these organisms the dominant form of TFBS discovery is computational estimations using techniques such as expectation maximization (EM). The goal of this work is to characterize the PR of ABA signaling pathway (ABAsp) genes using an in silico approach. To perform our analyzes, we used expression data of the sugarcane variety RB83-5486 to select all ABAsp genes differentially expressed in drought-tolerant plants. The 22 selected genes were mapped on the sugarcane genome SP80-3280 and the regions of 2000 nucleotides usptream from the transcription start site of each gene were extracted as putative promoter regions. For these regions we tried 3 different motif-finding approaches: Gibbs Sampling (using GLAM2), position-specific score matrices for previously characterized motifs in the JASPAR plant databases, and expectation maximization (using MEME). Only the last approach resulted in consistent results. GLAM2 showed a bias for AT-rich motifs and none of the results had any TOMTOM match against the JASPAR plant databases. Analysis using PSSMs did not find any candidates with significant scores. We parametrized MEME to find motifs from 5 to 15 nucleotides and maximum of 6 different motifs. We only considered motifs with a TOMTOM match to JASPAR plant databases. In general, 50% of the sequences presented similar TFBS architectures, with the bZIP, WRKY and AP2 / ERF classes of TFBSs as the most representative. Moreover, we distinguished different architectures for up and for down-regulated genes: in up-regulated genes we found motifs associated to ARR and bHLH TFBS classes, and in down-regulated genes were found motifs associated tot he HD-Zip, MYB and NAC TFBS classes. In this scenario it is possible to infer that the drought tolerance may be due to the crossing of different signaling pathways for water stress. Since two groups of TFBS distinct from the ABAsp were identified in the promoter region, one associated with up-regulated genes, and one associated with down-regulated genes, the architecture of the promoter region may be the factor necessary to activate the drought tolerant character observed in the evaluated plants.

Funding: FAPESP and Capes

CAATINGA SOIL MICROBIOME: an ecological and biotechnological overview revealed by omics approaches

Melline Fontes Noronha[1], Gileno Vieira Lacerda Junior[2], Renan Abib Pastore[2], Valéria Maia de Oliveira[2]

1 CENTRO PLURIDISCIPLINAR DE PESQUISAS QUÍMICAS BIOLÓGICAS E AGRÍCOLAS , UNICAMP
2 MICROBIAL RESOURCES DIVISION, RESEARCH CENTER FOR CHEMISTRY, BIOLOGY AND AGRICULTURE, UNICAMP

Abstract

Caatinga is a biome unique to Brazil characterized by two well-defined seasons, rainy and dry, mainly driven by the sparse rainfalls. Anthropogenic processes and climate change have caused worrying environmental damage, such as accelerated desertification and impacts on biogeochemical processes in this biome. Although some studies have shown that Caatinga soil microbiome is shaped by seasonal variation, it is still not clear what metabolic features allow microbes to adapt to environmental changes, in addition to the underlying effects over biogeochemical cycles. Recently, Caatinga has been the focus of intense research due the large input of vegetal organic matter from the falling leaves, which could recruit multiple microbial species producing a variety of plant biomass-degrading enzymes and providing a unique genetic resource for mining enzymes used in biofuel production. In this study, we aimed to explore the Caatinga soil microbiome under both ecological and biotechnological perspectives. The use of metagenomics to unravel the ecological landscape of the semiarid Caatinga soils showed that the microbial structure of pristine soil was shaped primarily by seasonality, with a strong increase of Actinobacteria and Proteobacteria members in the dry and rainy seasons, respectively. In contrast, Proteobacteria and Acidobacteria were notably altered by soil chemical parameters that played a critical role in shaping the microbial community from irrigation and fertilization-affected soils. Functional annotation identified a broad range of cellular processes related to osmotic and oxidative stress responses in microbial communities of pristine soils under seasonal variation. Metatranscriptomics analysis of irrigation-affected soils revealed a high contribution of Actinobacteria and Bacilli in the microbial community structure during the rainy and dry season, respectively. A higher abundance of transcripts for central carbohydrates metabolism, CO2 fixation and monosaccharide subsystems were found in the dry season (irrigation-affected). On the other hand, oxidative stress and protein degradation subsystems were overrepresented in the rainy season soils. Under the biotechnological perspective, results from data mining of Caatinga soil metagenomic libraries have shown a wide repertoire of lignocellulose degrading genes from a wide range of bacteria that could be explored for biofuels application. Although poorly studied, Caatinga soils is considered a promising source for the understanding of the microbial mechanisms for survival in harsh environments as well as for biotechnological applications.

Funding: FAPESP

Bioinformatic Analysis of Ubiquitin-Specific Protease Genes in Genome of Phaseoulus vulgaris L.

Monize Angela de Andrade[1], Daniel Alexandre Azevedo[1], Laurence Rodrigues do Amaral[1], Felipe Teles Barbosa[1], Enyara Rezende Morais[1], Matheus de Souza Gomes[1]

1 UFMG

Abstract

The Bean is a leguminous plant with high protein value, nutritional and heme iron donor widely consumed. The ubiquitin-proteasome is a pathway responsible controls many cellular processes able controlled such as degraded of proteins flawed, with error of synthesis, and that are no longer necessary. Once they were marked with ubiquitin protein, they are degraded by the protein complex proteasome 26S. The complex ubiquitin-proteasome regulation is one mechanism of control post- translational regulatory of many proteins important, but also able to be controlled by other proteins, which are called Deubiquitinating enzymes DUB and their function control the ubiquitin binding, off and clear the programmed degradation. The DUBs are composed by five super families of proteins such as JAMMs (metaloproteases), Ubiquitin C-terminal Hydrolases - UCHs, Machado-Joseph Domain MJD, Ovarian Tumor Proteases - OTU and Ubiquitin-Specific Proteases-USPs (UBPs in plants). The UBPs are specific proteins which degraded ubiquitin and therefore the study of these proteins is very important for understanding the regulation of many cellular functions and physiological in plants. Thus, the aim of this study, was to identify, annotate, characterize and classify putative UBP proteins in the genome of Phaseolus Vulgaris L.. Genome sequence of Phaseolus Vulgaris L. deposited in the public database Phytozome was used as queries in BLAST tool (Basic Local Search Alignment Tool). Conserved domains, amino acid residues from active sites were retrieved through the predicted proteins using PFAM database (http://pfam.sanger.ac.uk/) and CDD. Phylogenetic analysis was conducted in Mega5.2 program. We found 15 putative proteins UBPs in P. vulgaris among 12 subfamilies: UBP2-like; UBP4-like; UBP6-like; UBP8, UBP9-like; UBP13-like; UBP15, UBP17 and UBP18-like; UBP20-like; UBP22-like; UBP23-like; UBP25-like; UBP26-like. The putative UBP proteins showed conserved domains UCH containing significant and conserved residues at critical positions on the protein (putative active sites). The putative conserved catalytic site comprised (C/D/H) which divided into cys box and his box. The putative proteins UBP clustered on the phylogenetic tree in distinct clades agreeing with the predicted paralogous sub-families. Therefore, this study expanded the knowledge of the Ubiquitin-specific protease in P. vulgaris and it is the starting point for new challenges that pathway can help for produces, in future, cultivars genetically modified, with best growing, adaptation and production.

Funding: FAPEMIG, CNPq, UFU and CAPES

Gene Assembly, Prediction and Phylogenomic Analysis of Erianthus arundinaceus, a crop for biomass production

Nicholas Vinicius Silva[1], Luciana Souto Mofatto[1], Juliana José[1], Gonçalo Amarante Guimarães Pereira[2], Marcelo Falsarella Carazzolle[3]

1 UNICAMP
2 BRAZILIAN BIOETHANOL SCIENCE AND TECHNOLOGY LABORATORY, BRAZILIAN CENTER FOR RESEARCH IN ENERGY AND MATERIALS, BIOLOGY INSTITUTE - UNICAMP
3 BIOLOGY INSTITUTE - UNICAMP, NATIONAL CENTER FOR HIGH PERFORMANCE COMPUTING

Abstract

Erianthus arundinaceus is a wild perennial C4 grass, considered closely related to Saccharum. It has a good perennial ratooning ability, excellent vigor, high fiber and low sugar content, waterlogging and diseases resistance. Due to its high biomass production and strong tolerance to environmental stresses, it is regarded as one of the most promising crops for biomass production and source of desirable traits genes for breeding programs in sugarcane. The aim of this research was providing genomic information and phylogenomic analysis of Erianthus arundinaceus, in order to understand the evolutionary relationships among this species and other crops. Genomic sequences were obtained through Illumina MiSeq platform, generating 93 million of paired-end reads. The sequences were assembled using "The Polyploid Gene Assembler (PGA) Pipeline" with reference-based genomes of Sorghum bicolor, Zea mays, Setaria italica and Panicum virgatum, resulting in 15.596, 3.610, 2.532, 1.788 assembled sequences respectively. The remaining unmapped reads were De novo assembled using TRINITY, resulting in 389.960 sequences (N50=1923bp, Larger sequence of 15.566b. All sequences were used as reference for RNA-seq reads mapping using STAR, for the gene prediction analysis. Intron-exon junctions, provided by RNA-seq mapping, were used as hints for gene prediction in Genemark, resulting in 13.053 putative genes. These genes were filtered to find the most reliable, according to Blastp similarities with proteins from related genus. The 3.016 final reliable genes were used as training and test groups in AUGUSTUS, and gene predictions were performed with and without hints. The phylogenomic analysis among E. arundinaceus and five publicly available grasses genomes with reliable protein predictions (S. bicolor, S. bicolor "Rio", Z. mays, S. italica and B. distachyon) used the concatenated protein alignments of 472 single copy-ortholog genes identified with OrthoFinder. Proteins were globally aligned using T-COFEE, and submitted to Maximum Likelihood (ML) and Bayesian Inference (BI) phylogenetic analysis. We provide the first phylogeny with a genome dataset for E. arundinaceu, with strong branch supports and evidences that this crop is closely related to S. bicolor than previously inferred in literature. The new informations we present are central for further genome investigations and gene prospection from E. arundinaceus, and also for the development of new techniques for the improvement of sugarcane breeding in the production of biofuels and bioproducts.

Genomic analysis of Corynebacterium pseudotuberculosis strain 262

Raquel Enma Hurtado Castillo[1], Marcus Vinicius Canário Viana[1], Anne Cybelle Pinto Gomide[1], Vasco A. de C. Azevedo[1], Rommel Thiago Jucá Ramos[2], Artur Silva[3]

1 UFMG
2 UNIVSERSIDADE FEDERAL DO PARA
3 UFPA

Abstract

Corynebacterium pseudotuberculosis is a Gram-positive and facultative intracellular pathogen, causing important economic losses mainly in the ruminant production. The biovar ovis is nitrate negative and causes caseous lymphadenitis in sheep and goats, while biovar equi is nitrate positive and causes ulcerative lymphangitis, mastitis, and oedematous skin disease in a wide range of hosts. C. pseudotuberculosis 262 is an equi biovar strain isolated from cow milk. Genomic and phylogenomic analysis of C. pseudotuberculosis strains have been shown this strain as the most external of equi genomes and the closest one to ovis. In order to better characterize its genomic features, we present here a comparative genomic analysis between strain 262 and other 52 strains of C. pseudotuberculosis. A phylogenetic analysis based on a gene presence-absence matrix among strains of C. pseudotuberculosis does not cluster strain 262 in any of the two different clusters (biovars equi and ovis). Accessory genes shared between strain 262 and equi strains were predicted, such as a toxin and antirepressor, immunity-specific protein, CAAX protease self-immunity, serine hydrolase and superoxide dismutase. Accessory genes among 262 and ovis strains are genes related to ABC transporter protein, spermidine synthase, secreted protein and surface-anchored membrane protein. Sixteen genomic islands were predicted. Part of an island PiCp1 is a region of 10 Kb shared only with biovar equi, which presents CRISPR-associated protein. We also identified regions shared only with biovar ovis strains. In addition, strain 262 presented unique regions, containing 49 genes, such as MFS transporter, secreted protein, and hypothetical proteins, that could carry genes potentially associated with virulence. Finally, the pan-genomic analysis report accessory and unique genes that allow characterization of the strain 262. These findings enable us to better characterize the unique genomic features of strain 262 and generate new hypothesis to understand the differentiation of the C. pseudotuberculosis.

Funding: CNPq, CAPES

16S rRNA GENE-BASED PROFILING OF HOWLER MONKEY FECAL MICROBIOTA

Raquel Riyuzo de Almeida Franco[1], Júlio César O. Franco[2], João Carlos Setubal[1], Aline Maria da Silva[1]

1 USP
2 UNIFESP

Abstract

Howler monkeys (Alouatta spp.) are endemic animals from the Atlantic Forest biome that can be found in primary and secondary forests and even in small forest fragments. Their diet is based on tree leaves and fruits, depending on the season. This study aims to investigate the diversity of gastrointestinal bacterial community from captive and non-captive howler monkeys that inhabit São Paulo Zoo Park to correlate possible differences between their respective microbiotas and diets. We have collected a total of 25 fecal samples from captive and non-captive individuals at different seasons in 2013-2015. Total DNA extracted from the samples were then analyzed by 16S rRNA gene V3-V4 amplicon sequencing using the MiSeq-Illumina platform. In addition, a 16S amplicon sequence dataset of fecal samples from Mexican black howler monkeys was incorporated in the analyses. The sequences were used for alpha- and beta-diversity estimates, as well as for phylogenetic profiling using mostly the QIIME package. Our initial results point to differences both in the microbial community profile and diversity between captive and non-captive groups. When the microbial composition present in fecal samples of Brazilian monkeys were compared with Mexican monkeys, we observed that the microbial community of Brazilian captive individuals are very different from the other groups. Among the identified genera, we observed higher abundance of Bacteroides and Prevotella in the microbiota of captive animals. In humans, these two genera have been related to diets high in fat/protein and carbohydrates/fiber, respectively.

Funding: FAPESP, CNPQ, CAPES

A graph-based approach to explore local structures in genome graphs aiming the identification of genetic variants

Rodrigo Theodoro Rocha[1], Georgios Joannis Pappas Junior[1]

1 DEPARTAMENTO DE BIOLOGIA CELULAR, INSTITUTO DE CIÊNCIAS BIOLÓGICAS, UNB

Abstract

Advances in DNA sequencing technologies charter the possibility to generate hundreds of genomes to capture the individual genetic variability within a population. Currently, in order to compare these, a reference genome is used as a proxy to contrast the individuals. However, a single reference falls short in representing the wealth of genetic variation. In recent years, new graph based data structures were developed to capture sequence polymorphisms along a set of genomes, collectively named genome graphs. The rationale is to reduce bias and improve biological inferences from a one-dimensional universal reference (i.e. linear sequence representation) to a multi-dimensional (i.e. genome graphs) representation of multiple genomes. This is not a trivial change of perspective and challenges common tasks in bioinformatics, including read mapping and variant detection. Borrowing concepts from graph theory, specially those concerned with connectivity of graphs, we developed a framework to identify local graph structures (motifs) that represent sequence variability sites in genome graphs. These motifs can be simple genetic variants (i.e. bubbles) or more complex structures (i.e. super-bubbles). The strategy is based on the genome graph decomposition into biconnected components, and those into triconnected components that have specific characteristics. A graph is said to be biconnected (triconnected) if it has no set of 1-vertices (2-vertices) whose removal increases the number of connected components (i.e. splits the graph). We aimed at decomposing a genome graph with respect to its triconnected components with the aid of a data structure named SPQR-tree. This can be done in linear time and we show that some graph motifs are embedded in a subset of nodes of the SPQR-tree permitting the identification of genetic variants, whereas the paths connecting the nodes therein entail the allelic variants. Moreover, given that we can sort the identified graph motifs by complexity, it is possible to correlate these to hotspots of genetic variation in the genome graph. We expect that this approach will help to interrogate pangenomes to collectively identify hyper variable sites or genomic regions under selective pressure.

Funding: Programa de Pós-graduação em Biologia Molecular, UnB e CAPES

An NGS approach to analysing HMF resistance in Saccharomyces cerevisiae

Lucas Miranda[1], Sheila Tiemi Nagamatsu[2], Fellipe Melo[3], Bruna Tatsue[4], Gonçalo Amarante Guimarães Pereira[5], Gleidson Silva Teixeira[6], Marcelo Falsarella Carazzolle[7]

1 UNIVERSIDAD DE BUENOS AIRES - DEPARTAMENTO DE QUÍMICA BIOLÓGICA
2 BRAZILIAN BIOETHANOL SCIENCE AND TECHNOLOGY LABORATORY, BRAZILIAN CENTER FOR RESEARCH IN ENERGY AND MATERIALS CNPEM, BIOLOGY INSTITUTE - UNICAMP
3 BIOLOGY INSTITUTE ; UNICAMP
4 BIOLOGY INSTITUTE - UNICAMP
5 BRAZILIAN BIOETHANOL SCIENCE AND TECHNOLOGY LABORATORY, BRAZILIAN CENTER FOR RESEARCH IN ENERGY AND MATERIALS, BIOLOGY INSTITUTE - UNICAMP
6 BIOLOGY INSTITUTE - UNICAMP, FACULTY OF FOOD ENGINEERING - UNICAMP
7 BIOLOGY INSTITUTE - UNICAMP, NATIONAL CENTER FOR HIGH PERFORMANCE COMPUTING/UNICAMP

Abstract

Bioethanol is the most promising renewable fuel to substitute fossil fuel and it is generated as a product of fermentation of sugars, which can occur throught two processes, first generation (1G) and second generation (2G). They differ basically in the raw material used, where, in brazilian production, 1G require sugarcane juice, while 2G, unused plant parts (bagasse), which are rich in polymers such as lignin, cellulose and hemicellulose. Both methodologies involve subproducts that interfere in the metabolism of microorganisms used to fermentation step, being Saccharomyces cerevisiae the most commonly used. While first generation ethanol exhibits factors such as temperature, O2 pressure, pH, alcohol concentration and contaminants as inhibitors, second generation, for requesting a thermic preprocessing to expose the fibres, and to convert cellulose and hemicellulose to simple monomers, shows, besides 1G inhibitors, acetic acid, furfural and HMF - hydroxymethylfurfural. In summary it introduces the importance of understanding yeast's resistance to increase productivity. In this work we studied a diploid industrial strain resistant to HMF that was sporulated, beeing selected four resistent spores and three non resistent spores. These eight strains were sequenced using Illumina paired-end technology and the data was analysed. The pipeline includes quality analysis of the reads (FastQC), genome assembly of a non resistant strain (SPAdes), gene prediction (Augustus), variant calling (GATK) and effect prediction (VEP), gene annotation (Blastn), chromosomal mapping (MUMmer package) and gene ontology classification (SGD website). This pipeline allowed us to identify a set of 68 candidate genes that could be related to HMF robustness. Future perspectives include the application of this pipeline to a greater set of sequenced strains,

Respiratory nitrate reductase metabolic pathway in Corynebacterium pseudotuberculosis biovar Equi

Sintia Almeida[1], Vasco A de C Azevedo[2]

1 USP
2 UFMG

Abstract

Corynebacterium pseudotuberculosis can be classified in two biovars, based on their ability to convert nitrate to nitrite. The nitrate-positive biovar is Equi, which causes ulcerative lymphangitis in equines, while the nitrate-negative biovar is known as Ovis, which is the etiologic agent of caseous lymphadenitis in small ruminants. Both diseases are globally distributed and cause large economic losses to goat, sheep, horse and cattle farmers. The nitrate reduction is associated with the bacterium's ability to breathe in the absence of oxygen and having two different metabolic pathways, (1) respiratory nitrate reductase and (2) dissimilatory nitrate reduction. In the first pathway, the denitrification process takes place where the nitrate is sequentially reduced to nitrite, nitric oxide, nitrous oxide, and finally to dinitrogen. In the second pathway nitrate is directed converted into ammonia, which is secreted from the cell, this process can be performed by organisms with the nrf gene. This is a less common method of nitrate reduction than denitrification in most ecosystems. Prokaryotic nitrate reductases include a class of assimilatory enzymes and two classes of respiratory enzymes, all contain a guanylate molybdenum cofactor, but differ in their substructures, cellular location, and requirement for cofactor. Variability among enzyme is also found into the classes. Aiming to discover the molecular mechanisms to related the ability bacteria nitrate reduction, 19 complete genomes of C. pseudotuberculosis were analysed. To identify the nitrate pathways, genes of these pathways were analyzed using databases such as BioCyc , ENZYME , Keeg . For the analysis of metabolic pathways Pathway tools tools were used. Were done in the blast database Uniprot and protein domain analysis through INTERPROSCAN. Genome analysis revealed that C. pseudotuberculosis biovar Equi possess narKGHJI gene clusters that are similar to the narK gene and narGHJI operon of Escherichia coli. The gene encodes a nitrate/nitrite transporter, whereas the operon encodes a respiratory nitrate reductase (NarGHI) and one specific chaperone (NarJ) required for insertion of Mo-bisMGD cofactor in NarG. The enzymes that are involved in electron transport chain are also identified by in silico methods. Findings about pathogen metabolism can contribute to the identification of relationship between nitrate reductase and the C. pseudotuberculosis pathogenicity, virulence factors and discovery of drug targets.

Funding: CNPQ, FAPEMIG

Microbial diversity of inocula and mature compost from thermophilic composting operation at the São Paulo Zoo

Suzana Eiko Sato Guima[1], Laís Uchôa Rabelo Mendes[1], Roberta Verciano Pereira[1], Layla Martins[2], Aline Maria da Silva[3], João Carlos Setubal[3]

1 USP

Abstract

Waste composting harbors a high diversity of microorganisms that participate in organic matter degradation. Some of these microorganisms can be promising thermophilic candidates able to degrade plant biomass and produce biorefinery matter. Motivated by high microbial diversity in plant waste composting, our goal is to analyze microbial diversity in inocula and mature compost of the composting process operated by the São Paulo Zoo. These inocula were collected from compost pile in later phases, usually just prior to a turning procedure. These samples are collected because practice has shown that, when added to the waste material at the start of composting, they speed up organic matter degradation. Mature compost refers to the 100-day composting material ready to be used as biofertilizer. This final matter is likely to comprise microorganisms selected by low availability of nutrients and composting thermophilic conditions. In order to investigate the microbial composition and abundance in inocula and mature compost, we collected six inoculum samples from different compost piles and three mature compost samples. Amplified 16S rRNA genes were sequenced on Illumina MiSeq platform. We used USEARCH for OTU clustering, RDP classifier for taxonomy assignment and QIIME (Quantitative Insights in Microbial Ecology) for diversity analysis. After analyzing microbial abundance, we compared our samples with São Paulo Zoo composting time-series samples from a previous study. Taxonomic profile for initial inoculum exhibited reasonable similarity with composting final stage. The most abundant phyla for inoculum and mature composts were Firmicutes, Proteobacteria, and Actinobacteria. For inoculum, Actinomycetales, Clostridiales and Bacillales were the most abundant orders. Actinomycetales and Clostridiales were also the most abundant for mature composts, but Bacillales was present in less abundance. Shared OTUs between inoculum and time-series samples (ZC4) were higher on day 30 and on day 99 compared to other days. The five most abundant OTUs in mature composts were present in inoculum samples. Comparison between inoculum and time-series samples exhibited a trend in microbial dynamic and structure of the composting process with succession of some bacteria over others.

Funding: FAPESP, CAPES, CNPq

Oncogenic Fusion Gene CD74-NRG1 Confers Cancer Stem Cell-like Properties in Lung Cancer through a IGF2 Autocrine/Paracrine Circuit.

Takahiko Murayama[1], Tatsunori Nishimura[2], Kana Tominaga[1], Asuka Nakata[2], Noriko Gotoh[2]

1 THE UNIVERSITY OF TOKYO
2 KANAZAWA UNIVERSITY

Abstract

The CD74-Neuregulin1 (NRG1) fusion gene was recently identified in invasive mucinous adenocarcinoma of the lung, a malignant type of the lung adenocarcinomas, and is considered to be a new driver gene aberration. However, pathogenic functions of the CD74-NRG1 fusion gene have not yet been understood, and it is unknown whether the driver gene contributes to cancer stem cells (CSCs). Here, we show that expression of the CD74-NRG1 fusion gene induces CSC properties. Expression of CD74-NRG1 facilitated sphere formation of not only cancer cells but also non-cancerous lung epithelial cells. By using a limiting dilution assay in xenograft model, we showed that expression of CD74-NRG1 fusion gene enhances tumor initiating ability. We found that expression of CD74-NRG1 stimulates phosphorylation of ErbB2/3 and activates phosphatidyl inositol 3-kinase (PI3K)/Akt/NF-?B signaling pathway. Furthermore, we found that secreted insulin-like growth factor 2 (IGF2) levels were increased and phosphorylation levels of IGF1 receptor (IGF1R), the receptor for IGF2, were enhanced in a NF-?B activity dependent manner in cells expressing CD74-NRG1. These findings suggest that the CD74-NRG1-stimulated NF-?B activity induces IGF2 autocrine/paracrine loop. In addition, CD74-NRG1 fusion gene-induced tumor sphere formation was suppressed by treatment with inhibitors of ErbB2, PI3K or NF-?B or anti-IGF2 antibody. We thus provide evidence that the CD74-NRG1 fusion protein may not only act as a driver for tumor development but also initiate and maintain CSCs. Inhibition of ErbB/PI3K/Akt/NF-?B signaling pathway or secreted IGF2 are promising therapeutic strategies for CD74-NRG1 fusion positive cancers.

Funding: no

A predictive alignment-free model based on a new logistic regression-based method for feature selection in complete and partial sequences of Senecavirus A

Tatiana Flávia Pinheiro de Oliveira[1], Marcos Augusto Dos Santos[2], Marcelo Fernandes Camargos[3], Antônio Augusto Fonseca Júnior[3], Aristóteles Góes-neto[4], Edel Figueiredo Barbosa Stancioli[4]

1 DEPARTAMENTO DE MICROBIOLOGIA, INSTITUTO DE CIÊNCIAS BIOLÓGICAS, UFMG, MINISTÉRIO DA AGRICULTURA, PECUÁRIA E ABASTECIMENTO
2 DEPARTAMENTO DA CIÊNCIA DA COMPUTAÇÃO, INSTITUTO DE CIÊNCIAS EXATAS, UFMG
3 MINISTÉRIO DA AGRICULTURA, PECUÁRIA E ABASTECIMENTO
4 DEPARTAMENTO DE MICROBIOLOGIA, INSTITUTO DE CIÊNCIAS BIOLÓGICAS, UFMG

Abstract

In 2015, there was an outbreak involving pig farms in six Brazilian states, whose single agent found and described for the first time in the country was the Senecavirus A (SVA), a virus belonging to the genus Senecavirus (Picornavirales, Picornaviridae). This viral family also houses the genus Aphtovirus, whose species type is the Foot-and-mouth disease virus (FMDV), agent of Foot-and-mouth disease, a highly infectious disease notifiable under the strict control of the Ministry of Agriculture, Livestock and Supply (MAPA) and the World Organisation for Animal Health (OIE). In the past few years, there has been a growing interest in the application of methods of linear algebra and statistics in data mining, social networks, machine learning, bioinformatics and information retrieval. Among these methods, logistic regression approach draws some special interest as it is a standard method for data classification using genome data and is the most frequently used method for disease prediction. We introduce a model that represents sequences as 6-nucletotide frequency vectors in R4096 and 3- amino acids frequency vectors in R800 and uses information of SVA and FMDV from the complete genome or amino acid sequences of the polyprotein of theses viruses. In addition, partial sequences of nucleotides / amino acids of structural proteins (VP1, VP2, VP3 and VP4) of these viruses were used to build a new logistic regression-based method for classification. This new model allowed the assignment of values to parameters ai* that are associated with the frequency of a certain hexanucleotide or triplets codons of amino acids. Scrutinizing these parameters ai* unveiled that the most positive value may be related to important target sites of key virus proteins. Thus, this methodology was able to predict key regions in Senecavirus A, which can be important in studies of viral replication mechanism or in the development of diagnostic kits.

Funding: LANAGRO-MG ; FAPEMIG; CNPq

Acylsugar pathway in Solanum lycopersicum and Solanum pennellii

Thaís Cunha de Sousa Cardoso[1], Carolina Milagres Caneschi[1], Fernandes-brum C. N.[1], Matheus Martins Daude[1], Gabriel Lasmar Dos Reis[1], Lima A. A[1], Luiz Antônio Augusto Gomes[1], Laurence Rodrigues do Amaral[1], Chalfun-junior A.[1], Wilson Roberto Maluf[1]

1 UFU

Abstract

The cultivated tomato, Solanum lycopersicum, is one of the most important vegetable crops in global food and, together with the wild tomato Solanum pennellii are species widely used in developing cultivars. Although much is known about the S. lycopersicum and S. pennellii biology, little is known about the genes expression regulation involved in plant development and tolerance to biotic and abiotic stresses. Different allelochemicals present in wild tomato species were associated with resistance to pests, such the acylsugar. Acylsugars are fatty acids esters with 4-12 carbon atoms, containing glucose or sucrose and may act to reduce larval development, impairing feeding and oviposition of many tomato pests. The diversity of acylsugar produced in the tomato probably involves many genes and metabolic pathways involved in the acylsugar pathway. Thus, the study aimed to identify using in silico analysis and analyze the expression of the genes involved in the acylsugar metabolism of S. lycopersicum and S. pennellii. For the identification of the genes involved in the acylsugar pathway we used an optimized algorithm, BLAST tools and reference genes available in the NCBI, Phytozome v11.0 and SolGenomics. For expression analysis, we used three different tomato accessions: LA-716, TOM-684 and TOM-687 and leaves collected in three time stages (30, 60 and 90 days). It was extracted the total RNA and quantified in Nanodrop© ND-1000 to A260. For the expression analysis, we used the ABI PRISM 7500 Real-Time PCR, using SYBR Green and the cDNA obtained from the extracted RNA. The data was stored in 7500 Fast Software program. The reference genes used were eEF-1 and GAPDH. We identified 81 putative proteins in S. lycopersicum and 78 in S. pennellii involved in the acylsugar pathway, such as BCKD E2, IPMSA and TD, being key proteins in the pathway. There was a differential expression of BCKD E2 among the strains at all times. IPMSA showed a high expression in LA-716 access in the three times and TOM-684 and TOM-687 expressed an IPMSA level very low. TD showed different expressions between the ages of the plants in TOM-684 and TOM-687. Given the important role of the allelochemicals produced in Solanaceae, the results showed contribution to a better understanding of acylsugars, their processing pathway and their relationship with biotic resistance in S. lycopersicum and S. pennellii.

Funding: FAPEMIG, CNPq, UFU and CAPES

Alignment of the SSR microsatellite markers sequences with the cassava genome (Manihot esculenta)

Vanesca Priscila Camargo Rocha[1], Daniel Longhi Fernandes Pedro[2]

1 UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
2 UTFPR - PPGBIOINFO

Abstract

The species Manihot esculenta Crantz is cultivated in 105 countries, mainly in places located at latitude 30 ^{o} N and 30 ^{o} S (Nassar and Ortiz 2007), presenting wide adaptability to dry environments and to soils with low fertility, being able to withstand drought for periods of time leaf canopy and loss of water due to transpiration (El-Sharkawy 2004). The main product of cassava are the tuberous roots that are intended for human consumption, animal and in the industry of starch and flour processing. The objective of the present study was to perform the alignment of the microsatellite primers on the cassava genome and to verify in which chromosomes and ORFs are located. A comparative global alignment analysis of the 25 GA and SSRY forward primers in the cassava genome made available for the public use of these data in Phytozome (Manihot esculenta, version 6.1) was carried out and later the coding regions were applied in the tool provided by N.C.B.I. (National Center for Biotechnology Information) called Protein BLAST. Of the 25 loci analyzed, only five loci (SSRY 85, SSRY 65, GA 134, GA 140 and GA 161) were not located in the genome. The hypothesis for this case would be because they are in regions that do not encode proteins. Although, an alignment of the SSRY 85 loci was performed on the Arabidopsis thaliana genome and the same sequence was localized to seven possible genes in the genome. Some loci obtained identity below 28% as is the case of locus GA 12 (CHR 16, ORF 1); Locus SSRY 100 (CHR9, ORF 3); Locus SSRY 19 (CHR 1, ORF 3 and CHR 10, ORF 1). The GA 136 locus obtained 85% identity that is linked to expression for protein kinase, SSRY 19 with 81% expressed the beta-amylase protein, SSRY 27 with 94% identity for ethylene receptor expression. Microsatellite molecular markers can be used to identify genes that correspond to proteins.

Funding: Fundação Araucária

An analytical pipeline for detection of differential DNA methylation from restriction reduced genomic representation: a pilot study in Eucalyptus.

Wendell Jacinto Pereira[1], Marilia de Castro Rodrigues Pappas[2], Dario Grattapaglia[3], Georgios Joannis Pappas Junior[1]

1 DEPARTAMENTO DE BIOLOGIA CELULAR, INSTITUTO DE CIÊNCIAS BIOLÓGICAS, UNB
2 EMBRAPA RECURSOS GENÉTICOS E BIOTECNOLOGIA
3 EMBRAPA RECURSOS GENÉTICOS E BIOTECNOLOGIA, UNIVERSIDADE CATÓLICA DE BRASÍLIA.

Abstract

Phenotypic plasticity, the ability to display a range of phenotypes as a function of variable environments, is a key feature in land plants. Nevertheless, knowledge of the extent and underlying mechanisms of phenotypic plasticity in response to abiotic stresses is still fragmentary. Besides genetic diversity, epigenetic variation is believed to contribute to tree phenotypic plasticity and adaptive potential. We set out a new approach to perform genome-wide differential methylation analysis by means of parallel construction of double digestion restriction libraries, namely PstI-MspI (methylation insensitive) and PstI-HpaII (methylation sensitive), followed by short-read NGS sequencing. For technical validation we evaluated the differences in methylation patterns in three tissues (xylem, juvenile and adult leaves) from clone BRASUZ1, the tree sequenced in the Eucalyptus grandis genome project. A new computational pipeline was created, using open source tools, to process this type of NGS data in a fully reproducible way. The final goal is to identify and annotate differentially methylated regions across samples with the assumption that the data follows a negative binomial distribution. The pipeline results for this experiment provided reproducible, genome-scale methylation measurements at 22,000 genomic sites consistent in all tissues and biological replicates. From this, 4,000 methylated sites were observed and the majority (64%) conserved among the analyzed tissues. On the other hand, approximately 6% of the sites were specific for each of the three tissues. Considering the genomic context, 58% (2,335) of the verified methylated sites fall within genes, whereas 570 (14%) are in transposons. Contrarily to the expected, the methylation profile in the genic space is favored and this experimental bias offers a cost effective alternative to contrast epigenetic states of a sizeable fraction of genes in plant genomes. At the end of the study, we expect to provide a flexible tool for easy execution of this approach to other species.

Funding: CNPq, EMBRAPA Recursos Genéticos e Biotecnologia, Programa de Pós-Graduação em Biologia Molecular - UnB.

Improving variant accuracy with Copy number variant pipeline for target sequencing

George de Vasconcelos Carvalho Neto[1], Wilder Barbosa Galvao[1], Marcel Caraciolo[1], Rodrigo Bertollo[1], Joao Bosco Oliveira[1]

1 GENOMIKA

Abstract

Studies comparing human genomes have been shown that more base pairs are altered as a result of structural variants (SVs), including copy number variants (CNVs), than as result of point mutations. Structural variants were first defined as insertions, deletions and inversions greater than 1 kb size. However, with the high-throughput sequencing becoming a routine for genome analysis, the spectrum size of SVs and CNVs have been extended to events >50 bp in length. Due to the cost and complexity of analyzing whole genome sequence data, target sequencing (TS) has become the major approach for clinical diagnostic purposes. Also, TS allows for the detection of CNVs in addition to single-nucleotide variants (SNVs) and small insertions/deletions (INDELs). More recently, novel tools have been developed to improve CNVs identification from targeted panel sequencing data. With efforts conducted to develop an internal protocol for CNVs detection in TS data, and increase possible genetic case elucidation, the aim of this study was to analyze the performance of state-of-the-art CNV detection methods on targeted next-generation sequencing (NGS) before implementation at our laboratory. The chosen softwares to the analyses were ExomeDepth, panelcn.MOPS, CoNVaDING and VisCap. Based on samples from patients that sequenced BRCA1 gene and had variant confirmed by Multiple ligation probe assay (MLPA), which represents the gold standard for molecular analysis of diseases caused by CNV, is possible to evaluate accuracy and sensitivity for each tool. Finally, in the end of this evaluation we expect to maximize our variant detection accuracy using the best algorithm tested or the combination of the best callers for CNV on target sequencing data.

Funding: Genomika Diagnósticos

Best Practices for Bioinformatics Pipelines for Molecular-Barcoded Targeted Sequencing

Marcel Caraciolo[1], Wilder Barbosa Galvao[1], George de Vasconcelos Carvalho Neto[1], Rodrigo Bertollo[1], Joao Bosco Oliveira[1]

1 GENOMIKA

Abstract

In cancer research the detection of mutations is critical, for tumor samples and blood samples mutations may be present in very low fractions of DNA molecules. By using molecular barcoding technology, more than reduce the impact of enrichment, the sequencing errors can be eliminated by tagging each input molecule with an unique molecular identifier (UMI). In contrast to sample barcoding, molecular barcoding assigns a unique sequence not just to all the molecules from a certain sample, but to all molecules being amplified and sequenced. Despite the difference it is common to have both sample barcodes and molecular barcodes in the same sequencing reads. Recent works on this approach show outstanding performance in targeted high-throughput sequencing, being the most promising approach for the accurate identification of rare variants in complex DNA samples, and has application in several areas such as detecting DNA mutations at very low allele fractions with high accuracy for cancer samples and reducing sequencing artifacts occurrences. However, at the sample preparation, the residual PCR errors might be introduced at first PCR cycles and during UMI tag attachment, which decrease the accuracy of variant calling. In order to perform the variant detection on those input data, a different approach is required for bioinformatics pipelines that handles the caveats of UMI-based analysis. By using specific algorithms and softwares, the pipeline is designed to obtain high-fidelity mutation profiles and call ultra-rare variants. In this poster we present the best practices and strategies for handling the UMI-tagged data, by showing the steps and related software tools to the audience when building the variant calling pipeline

Funding: Genomika

4 | Phylogeny and Evolution

An Approach to Study Taxonomic Distribution of Genes: Biofilm Production as a Model

Antonio Gilson Gomes Mesquita[1], Sabrina Sondre de Oliveira Reis Margarido[1], José Miguel Ortega[2], Tetsu Sakamoto[3]

1 UNIVERSIDADE FEDERAL DO ACRE, LABGENMOL
2 UFMG, LABORATÓRIO DE BIODADOS
3 UFMG. LABORATÓRIO DE BIODADOS

Abstract

Taxonomic distribution of the orthologous of a gene of interest is not a trivial task to accomplish. Manual inspection of phylogenetic trees consist in the most analytical procedure to investigate the lowest common ancestor (LCA) and the descendant clades that have or not the gene, since deletions and lateral gene transfers may occur. Here we present the analysis of the taxonomic distribution of genes involved in production of bacterial biofilm. The query genes have been selected from the protein database UniProt and the taxonomic distribution was analyzed with software TaxOnTree. Originally, the software run a BLAST search limited by e-value (1e-5), similarity threshold (e.g. 30%) and builds phylogenetic trees using MUSCLE to obtain a multiple alignment, TrimAL to edit the alignment, FastTree to build the phylogenetic tree and colorizes the branches according to any taxonomic rank of choice, e.g. phylum, class, order, etc. turning the user into a taxonomy expert. However, since our interest was on bacterial genes, the limit of 200 proteins to generate the trees often did not surpass the analysis of a genus, due to the presence of great number of taxonomically closely related orthologues. Thus we further developed TaxOnTree to be able to present a restrict number of hits from a chosen clade, e.g. phylum. Development of the tool to provide biologically meaning outputs was developed in parallel to the study of biofilm genes. We could map the clade of origin of a specific gene (LCA) and the relationship of it to more distant ones. The genes of interest were: (i) phaZ, almost all clusters contained bacteria from the order Burkholderiales, two clusters containing Chromobacterium violaceum, one with sequences from phylum Ascomycota, one with Actinobacteria, along with a cluster of animal and protozoa sequences; (ii) phaE, restricted do Cyanobacteria of distinct classes; (iii) phaB restricted to Proteobacteria from the class Betaproteobateria; (iv) phaZ1, showing a cluster with Chromobacterium violaceum and Cyanobacteria associated with Proteobacteria from order Neisseriales and other clusters from Actinobacteria and Ascomycota, and one with animal and protozoa sequences; (v) phaZ2, with the same distribution of phaZ1; and (vi) phaC from Chromobacterium violaceum and present in Actinobacteria, Cyanobacteria and also Porifera and Mollusca.

Funding: Universidade Federal de Minas Gerais. Laboratório de Biodados.

Evolution of Bitopic Signal Transduction Proteins

Aureliano Coelho Proença Guedes[1], Raphael D. Teixeira[1], Chuck S. Farah[1], Robson Francisco de Souza[1]

1 USP

Abstract

Sensing environmental changes and relaying this information to inside the cell is very important for bacteria and other organisms. Bitopic signal transduction proteins are a diverse array of proteins that possess both exposed extracellular sensor domains and cytoplasmic domains connected by transmembrane regions. Changes in environmental conditions or the presence of external chemical stimuli are detected by the extracellular domains of these proteins leading to structural changes at their cytoplasmic portions. Structural rearrangements of the cytoplasmic domains result in intracellular activities that affect cellular behavior, such as protein post-translational modifications or synthesis of small molecules that act as secondary messengers. Recent studies have shown that XAC2383, a class I of periplasmic-binding protein (PBP) protein of Xanthomonas citri, physically interacts with the periplasmic Cache domain of the bitopic GGDEF protein XAC2382, which is encoded by an adjacent downstream gene. Similar PBP domains were found to be encoded by genes located just upstream to other bitopic proteins, most of which are composed by distinct combinations of a periplasmic Cache and various cytoplasmic output domains, such as histidine kinases, methyltransferases, cyclic nucleotide synthases and Sigma54 activators. Using homology searches for PBPs and the analysis of domain architectures and genomic context, we demonstrate that distinct classes of PBP domains are often fused to output domains in proteins with varying levels of domain architectural complexity. Phylogenetic analysis of PBP's and Cache domains of bitopic proteins encoded by adjacent genes suggest that the more complex bitopic proteins originated from events of gene fusion and possibly gave origin to new, simpler architectures through loss of internal domains. Importantly, the same pattern was also observed for periplasmic-binding proteins of class II, thus suggesting that a common mode of evolution for new architectures, involving the fusion of adjacent genes and subsequent loss of internal protein domains, could be a general feature of the evolution of bitopic signal transduction proteins and also a recurrent event, that occurs independently in different lineages of sensor and output domains. In order to evaluate this hypothesis, we will now consolidate our search strategies into a pipeline for comparative genomics and protein evolution and extend our analysis to other combinations of bitopic proteins and a broader range of extracellular sensory domains. Our results will help us understand the relative impact of recombination and gene fusion on the evolution and shuffling of multidomain proteins.

Funding: CAPES, FAPESP

Systemic study of the evolution of flowers

Beatriz Moura Kfoury de Castro[1], Tetsu Sakamoto[2], Carlos Alberto Xavier Gonçalves[1], José Miguel Ortega[2]

1 UFMG
2 UFMG, LABORATÓRIO DE BIODADOS

Abstract

Flowers are recent innovations in the evolutionary history of plants on the geological timescale of plant diversification. They are the reproductive structures of angiosperms (flowering plants), which constitute the most diverse and cosmopolite group of plants. In order for flower to be formed, a complex gene regulatory networks control the floral development, involving several genes. The main purpose of this work was to collect the current knowledge about the molecular basis of floral development and assemble it in the format of a metabolic pathway. Since flowers are recent innovation among angiosperms, we also investigated the evolutionary origin of flowering genes, verifying if they have originated recently along with the appearance of flowering plants. To retrieve genes associated to flowering process, we collected 1000 scientific articles depicting the molecular biology of flowering. The list with articles of PubMed IDs (PMIDs) resulting from search were submitted to text-mining tools, to assist us on determining genes and biointeractions described on them. After this selection, construction of the metabolic pathway was done with manual curation. To analyze the evolutionary origin of those genes retrieved from the text-mining, their sequences were retrieved from Uniprot database, their orthologs from other species were retrieved using SeedServer and the Lowest Common Ancestor (LCA) was determined and used to infer the gene origin. Phylogenetic inference was also applied to analyze in more depth the evolution of some genes in the pathway. As result, we found 80 genes to be linked to flowering process. Among them, 20 belong to MADS-box family, demonstrating its relevance on shaping the flower. Gene origin inference indicated that genes comprising the floral development pathway appeared on different clades during the evolutionary history of the plants. However, some proteins of MADS-box family seem to have originated during the conquest of the terrestrial environment by the plants, well before the appearance of the flower structure. However, some genes were found exclusively on the clade of the angiosperms (Magnoliophyta) and non-basal angiosperms (Mesangiospermae). A more in-depth investigation of the evolution of MADS-box genes participating in flower development showed that recent and multiple duplications have taken place on this family among the flowering plants. Since the period of the expansion of this family coincides with the origin of angiosperm, it suggests that these events have contributed greatly to the appearance of the floral structure.

Funding: UFMG

11.000 Synonymous! But not so much...

Clovis Ferreira Dos Reis[1], Rodrigo Juliani Siqueira Dalmolin[1], Andre Fonseca[2], Sandro Jose de Souza[2]

1 UFRN

Abstract

Mutations that alter the amino acid sequence of a protein are known to be under natural selection while synonymous mutations are assumed to be neutral regarding protein function. Indeed, synonymous mutations have been used as a proxy for neutral alterations in genomes and as a reference against which potential selected mutations are being compared. However, it is becoming quite clear that at least a fraction of those synonymous mutations has deleterious effects. This work aims to create a method that correlates the influence of synonymous mutations, based on the genome codon bias over prokaryotes fitness. To perform such analysis was used data of the genome sequence of 12 E.coli populations sequenced 11 times over 50,000 generation (Tenaillon et al. 2016). To evaluate the putative impact of synonymous changes, we used the Relative Adaptiveness of a Codon (w) developed by Sharp (1987), in which every individual codon frequency is compared to an optimal codon frequency. Based on it, we proposed a w variation index (delta w) defined as w of mutation minus w of reference. A negative delta w would indicate that the new codon, resulting from the mutation, is less frequent in the genome of that species and likely associated to a less abundant tRNA. To evaluate whether synonymous mutations are randomly distributed regarding the delta w score, a 20,000 rounds Monte Carlo simulation was performed, in which the same number of real mutations was randomly created. As result, the pattern of Mont Carlo and real synonymous mutations set differs significantly and this discrepancy suggests that purifying selection is acting on synonymous mutations likely through tRNA abundance. Thus, the delta w seems to be an adequated index to evaluate the codon bias influence over microbial fitness and the results suggest a selection mechanism operating over synonymous mutations.

Funding: Universidade Federal do Rio Grande do Norte

THE ORIGIN OF THE GENES OF HUMAN DIGESTIVE SYSTEM SECRETION

Fenícia Brito[1], Tetsu Sakamoto[2], José Miguel Ortega[3]

1 UFMG
2 UFMG. LABORATÓRIO DE BIODADOS.
3 UFMG, LABORATÓRIO DE BIODADOS.

Abstract

The processes that ensure the maintenance of life require energy and for all heterotrophic organisms the source of energy are the nutrients available in the environment. Since the origin of the first forms of life the processes and systems involved in obtaining and metabolizing food have been determinant in the evolutionary success of the species. In humans the main stages of digestion depends on salivary, gastric, pancreatic and biliary secretions. Although studies of comparative anatomy in metazoa are well documented and the systemic knowledge of several processes is reported in many databases, studies addressing the comparative genomics and the evolution of the components of this system are scarce. Here our goal was to study the sequential origin of the genes involved in heterotrophy using the human digestive system as model. For this, software-recognizable diagrams for the digestive system secretion pathways were created based on models available in KEGG Pathway database. The origin of each component was estimated using tools for homologous clustering and for lowest common ancestor inference. This allowed us to infer the origin of the system based on the origin of its genes. Our results show that the most ancestral genes found in the pathways act on cell signaling processes and arose before the systems have been originated. The most recent components, such as receptors and some transporters, which are essential for secretion function, appeared from Metazoa. In addition, some components with auxiliary function, such as bicarbonate secretion in the pancreas and bile, have a more recent origin, indicating that this process appears as a refinement in these secretory pathways. Salivary secretion has the highest number of recent components and many of the proteins secreted are exclusive in mammals. In addition, we performed an analysis using the ELDOgraph program to identify which organisms or OTUs have more proteins close to human. The results show that the species that has more ELDO with human belong to the genus Pan (chimpanzees and bonobos). Analyzing at the order level the species with more ELDO with human belong to the orders Dermoptera and Rodentia. Our results show that the secretion pathways of the digestive system in mammals share many similarities, although some proteins are more distant. The data presented here allowed us to draw a scenario about the evolution of the digestive process, contributing to the evolutionary history of this system.

Funding: Programa de Pós-Graduação em Bioinformática

Comparative genomics of bacterial toxins associated with the type IV secretion system

Gianlucca Gonçalves Nicastro[1], Robson Francisco de Souza[2]

1 INSTITUTO DE CIÊNCIAS BIOMÉDICAS), DEPARTAMENTO DE MICROBIOLOGIA, USP
2 USP

Abstract

Bacteria are continuously exposed to biological conflicts, such as parasitism and competition. To succeed in such interactions, bacteria often deploy proteinaceous toxins that will kill other species through various mechanisms. Different secretory systems may be used to export toxins to the environment or directly into the target cells. Notably, the type IV secretion system (T4SS), previously known for its role in the translocation of virulence factors into eukaryotic cells and transport of genetic material, was recently recognized to participate in bacterial competition. However, in-silico prediction of proteins secreted by the T4SS is a difficult task, and currently available methods often fail to detect the new, experimentally verified, T4SS-associated toxins. In this work, we apply deep sequence similarity searches and genomic context methods to search for novel toxins that act as Tfes ("Type four secretion system associated effectors"). We found a novel T4SS-targeting signature that is conserved in the C-terminal part of proteins harboring N-terminal domains involved in DNA transport and a range of toxic enzymatic activities. Based on these features, we named this predicted domain as "T4SS C-terminal Tag 1" (T4CT1). T4CT1 is mainly found in Proteobacteria, but a few representative proteins are also present in other Gram-negative lineages, such as cyanobacteria. Importantly, genes that code for proteins with this signature are primarily located adjacent to T4SS loci. Our results show that comparative methods can be used to identify novel sequence signatures and, by iterating our approach, we intend to find new putative toxins secreted by this system. We expect that the detailed analysis and classification of the T4SS toxins will help us uncover the nature of the evolutionary processes influencing competition among different species of bacteria.

Funding: FAPESP, CAPES

Detection and recontruction of viral haplotypes from APMV-1 samples

Giovanni Marques de Castro[1], Francisco Pereira Lobo[1], Helena Lage Ferreira[2]

1 UFMG
2 USP

Abstract

Two samples of APMV-1 were sequenced using MiSeq and analyzed to verify which genotype the samples belong. As viruses have very high error rates when replicating it is expected to find variants given a high depth of sequencing as provided by NGS. The data generated for both samples were enough to analyze the underlying viral population, the quasispecies of APMV-1 for both samples. For this, the reads for each library were mapped to a few reference genomes and selected that which had the most mapped reads, the KJ123642. Using the aligned reads as the input for the software QuasiRecomb and restricting the region of to the F protein, the haplotypes for both samples were reconstructed. The most frequent haplotype from each sample and other 88 APMV-1 genomes from NCBI were used to reconstruct the phylogeny, the analysis of the phylogeny allow to visualize in which genotype they belong. The F protein is know to have a cleavage site in which the amino acid present can be used to predict if the virus is lentogenic or velogenic. The analysis of the cleavage site revealed that the most frequent haplotypes from both samples are velogenic. More than 65% of the reads were aligned to the reference genome (KJ123642) for each sample. The phylogenetic analysis showed that they group with the Vb genotype. Using more of the reconstructed haplotypes to reconstruct the phylogeny showed an extremely close result to using only the most abundant haplotype, clustering together, most likely due to the founder effect of a small related viral population.

Funding: Capes

Insights about the phylogenomic approaches to Staphylococcus aureus taxa clustering

Guilherme Coppini[1], Célio Dias Santos Júnior[1], Flávio Henrique Silva[1]

1 FEDERAL UNIVERSITY OF SÃO CARLOS

Abstract

Staphylococcus aureus is a widespread bacteria involved in resistance-acquired infections. Rapid evolving rates of S. aureus make approaches as 16S rRNA phylogenetic trees less usual, being the most usual method to identify S. aureus strains Staphylococcal Protein A (SpA) phylogeny, which is based on a single protein-coding gene with a hyper-variable region X. But, could the phylogeny of a gene does reflect a complex network of strain phylogenetic relationships? Our main goal was to compare the clustering resolution and accuracy of whole-genome distance based approaches and protein-based phylogeny inside a S. aureus dataset, correlating taxa clustering to find a better strain phylogeny method. To do this, 168 S. aureus genomes available in NCBI were selected and had their proteins predicted by PRODIGAL software. SpA sequences were selected using BEAF pipeline with a manually curated database, and 5 partial or disrupted ORFs were discarded. SpA sequences were aligned with MAFFT, had their substitution model predicted by ModelGenerator and best model was selected through Aikaike Informative Criterion. LG+G+F+I substitution model was used in RAxML program to generate the consensus final tree, following the extended majority rule from a bootstrap with 100 pseudo replicates. Genomes were fragmented into 250 bp fragments by a home-made python script and were pairwised searched through Usearch local alignment, generating a total of 26,569 alignments. Total sum of alignments' bit-scores for each possible genome pairs were calculated and then used to calculate euclidean distances, generating a dissimilarity matrix, used to generate the final dendrogram by UPGMA method. Different topological inferences were summarized as Compare2trees's score of 0.4034, where the topologies were almost 60% divergent. When compared using ETE toolkit, the two constructions regarded a normalized Robinson-Foulds coefficient of 0.94 and a symmetric distance (RF) of 279.0. A frequency of edges in the SpA tree of 0.57 was also found in genomic distance dendrogram. The distance results provided by Phylo.io showed a poor topological conservation between both approaches. These results showed a different topological resolution between both approaches, reflecting a different strains sorting. The analysis of strains relation carried using phenotype data revealed an apparent better resolution for the whole-genome distance approach, where most strains were correctly clustered, although more tests are needed to confirm it. In this sense, we observed the importance of considering wide-genome approaches for taxa clustering, which despite not necessarily reflecting the phylogeny, could still be used to reflect the phenotype.

Funding: CNPq, CAPES

Ancestral reconstruction of transthyretin / 5-hydroxy isourate hydrolase sequences

Lucas Carrijo de Oliveira[1], Laila Alves Nahum[2], Lucas Bleicher[1]

1 UFMG
2 CPQRR/FIOCRUZ MINAS

Abstract

Transthyretin (TTR) is a tetrameric protein - each of the four identical chains having about 130 amino acids (14 kDa) - that transport thyroid hormones in blood and brain. It also participates indirectly in transport of retinoic acid by coupling to retinol binding proteins. It was firstly described in eutherian mamals as a carrier of thyroxine (T4), however in most vertebrates it has more afinity to the active form of this hormone, triiodothyronine (T3). Mutations in this protein can lead to several diseases, like high concentrations of thyroid hormones in blood, or even formation of amiloid fibrils, associated to neurodegenerative diseases. Some evidence suggest that the gene for TTR has arised during the emergence of vertebrates, from a putative gene duplication of an enzime found since bacterias to vertebrates, involved in uric acid metabolism: the 5-hydroxy isourate hydrolase (HIUase). Since they are present in all kingdoms of life, have a stable and conserved structure, don't suffer post-translational modifications and are able to have their activity modified from enzime to hormone carrier by changing just a few amino acids in the active site, they are considered an excelent model for studies of molecular evolution, specificaly function divergence, in homologous protein families. One common way to assess protein evolution is to look at a multiple sequence alignment. More specificaly, some methods are capable of predicting putative ancestral sequences starting from nowadays sequences. In the present work, a bayesian phylogenetic analysis of sequences belonging to TTR/HIUase family was proceeded and, for some key nodes of the phylogenetic tree, ancestral sequences were reconstructed by maximum likelihood methods. Their 3D structure were predicted by similarity and their corresponding genes were submitted to synthesis for further experimental characterization. Some conservational patterns in active sites could be verified according to information available in literature, corroborating some hypothesis concerning specifity determinig positions.

Funding: CAPES, CNPq

Genome assembly completeness and its effect on phylogenetic estimation

Rafael Cabus Gantois[1], Raquel Enma Hurtado Castillo[1], Rodrigo Profeta Silveira Santos[1], Thiago de Jesus Sousa[1], Marcus Vinicius Canário Viana[1], Anne Cybelle Pinto Gomide[1], Artur Silva[2], Rafael Azevedo Baraúna[2], Vasco A de C Azevedo[1]

1 UFMG
2 UFPA

Abstract

Corynebacterium pseudotuberculosis is a Gram-positive bacteria that causes diseases in humans and animals around the world. It's divided in two biovars: Ovis Biovar infects goats, cattle and sheep and Equi Biovar infects equines and cattle. Currently, there are 73 genomes of this species available in NCBI database, 14 of these as drafts, and this number is still growing. Previously, the NCBI phylogenetic tree of this genomes had two clusters representing the two biovars. However, eleven draft genomes of Equi biovar were recently deposited and formed a third cluster, external to the previous ones, instead of being clustered within the other Equi. In this work, we are reassembly these draft genomes in order to investigate the effect of the genome assembly completeness on phylogenetic estimation. The genomes were sequenced at National Reference Laboratory for Aquatic Animal Diseases of Ministry of Fisheries and Aquaculture (AQUACEN) using Ion Torrent PGMTM platform and a 400 pb fragment library. The new assemblies are being performed using Newbler 2.9 by a de novo strategy, following drafting using CONTIGuator 2.7, and gap filling by reference assembly using CLC Genomics Workbench 7. The genome annotation is done automatically using RASTtk and manual curated. We will use PEPR to reconstruct two different phylogenomic trees with the C. pseudotuberculosis genomes available in NCBI: one using the drafts and other using the complete assembled genomes. As a preliminary result, the completely assembled genome of strain MB302 was clustered with the other complete genomes. As expected result, we hope that every single re-assembled genome will move to inside the tree too.

Funding: CNPQ, CAPES, FAPEMIG

Initial characterization of the blood DNA virome from 1000+ Brazilian elderly individuals

Suzana Andreoli Marques Ezquina[1], Michel Naslavsky[2], Maria Rita Passos-bueno[3], Mayana Zatz[3]

1 CENTRO DE ESTUDOS DO GENOMA HUMANO - CEGH -USP
2 CENTRO DE ESTUDOS DO GENOMA HUMANO - CEGH - USP

Abstract

The characterization of blood DNA virome is important for the identification of emerging pathogens in a population, and an interesting aspect of the integrated virus from infections and immunizations a person has in its lifetime, especially when we are studying people aging 60 years old and beyond. Humans harbor a huge number of endogenous retroviruses embedded in their genome, as remnants of an ancestral germline infection. These endogenous retroviruses may still contribute to pathological processes, including cancer. Whole genome sequencing (WGS) projects that involve whole blood DNA extraction allow the precipitation of external viruses along with the nuclear DNA, which are sequenced and not mapped to the human genome. These viruses contribute to the findings of current infections. Here we intend to present the initial results of the virome analysis, in particular, the determination of viral prevalence and distribution by sex and age. This cohort of 1324 individuals aging from 58 to 104 years (mean age74) is composed by a population-representative sample of São Paulo (SABE study, n=1199) and by a cognitively healthy octogenarians sample (n=125). SABE study individuals harbor the expected incidence of comorbidities under their age range. WGS was carried on the Illumina HiSeqX sequencer using 150 base paired-end single index reads, which were aligned with ISIS Analysis Software to the Human genome version hg38. Duplicate reads were removed. Samtools version 1.5 was used to extract the unmapped reads and the bam file was converted to fastq using bamToFastq and then to fasta using fastq_to_fasta. Blastn+ version 2.6.0 was used to find hits of the fasta files extracted from each individual with the NCBI database "RefSeq and Neighbor nucleotide records" with 116,503 entries of viral genomes. The Blast results were then parsed and counted for each type of virus in each individual using Perl scripts. We filtered putative viral matches from blastn+ using an e-value equal or less than 1e-10. The identification of viruses was performed in raw reads and in contigs after the assembly. In both cases the correlation between the number of reads/contigs and the hits on blast were comparable, but we noted that the assembly of reads in contigs using SOAPdenovo2 yielded overall better hits. Assembled contigs increased the number of longer matches and higher e-values, also decreasing the amount of time needed for the blast program to run. We obtained a mean of 8k hits in 350k queries. In this preliminary analysis, we found that abundances were very similar between women and men, although Human endogenous retrovirus and a strain of Human immunodeficiency virus 2 were slightly increased in women in contrast to Human herpesvirus and mammarenavirus which were slightly increased in men. Further investigation is necessary to better characterize the virome profile of Brazilians in comparison to different populations.

Funding: FUSP

Genome-wide identification of novel miRNAs in cnidarian genomes

Tamires Caixeta Alves[1], Laurence Rodrigues do Amaral[1], Matheus de Souza Gomes[1]

1 UFU

Abstract

Cnidarian is a phylum of the kingdom Animalia of great ecological and economic importance, due to its peculiar characteristics like, for example, its high adaptive capacity to the most diverse environments. The mechanisms that mediate changes in gene expression in response to stress remain unknown and there is a need to look at the regulatory mechanisms that control the dynamic expression of genes in the face of environmental challenges. Recent studies have shown the importance of gene regulation involving small RNAs, their processing system and their performance at the cellular level. MicroRNAs are considered one of the most important noncoding small RNAs silencing the mRNAs controlling their gene expression. Computational methods have been applied to identify and characterize putative miRNAs, their precursor and mRNA target genes. MiRNAs in plants and animals are already well elucidated. However, there is an evolutionary gap that needs to be addressed about the emergence of miRNAs. The most commonly accepted hypothesis is that of convergent evolution. We believed that studies of cnidarian can better explain it, either to affirm or deny the hypothesis. To obtain further clarification on the cnidarians, we searched for precursor, mature and miRNA targets in three species of cnidarians: Acropora digitifera, Exaiptasia pallida and Hydra vulgaris. An optimized algorithm, with stringent filters according to the conserved characteristics of miRNAs, was used to search for precursors. For search of mature miRNAs in the sequences of identified precursors, we aligned sequences deposited in miRBase version 21 (http://www.mirbase.org/). The RNAfold program was used to predict the secondary structure of the miRNA precursor and the RNAalifold and ClustalX 2.1 programs were used to generate alignments of the precursors and their orthologs. Phylogenetic analysis was performed with the aid of MEGA5.2 program. We identified 77 miRNA precursors, 77 mature, within 66 families in Acropora digitifera; 22 miRNA precursors, 22 mature, within 21 families in Exaiptasia pallida; 12 miRNA precursors, 12 mature, within 11 families in Hydra vulgaris. miRNAs plants and bilateral animals were found in our study. This suggested a common ancestor between cnidarians and plants. Phylogenetic analysis showed that our results corroborated the tree of life. Thus, our results expanded the study of small RNAs involved in cnidarians, and provided an alternative explanation on the evolution of miRNAs.

Funding: FAPEMIG, CNPq, UFU and CAPES

5 | Proteins and Proteomics

Structural and comparative analyses of fumarate hydratase from three species of Leishmania genus presented in Brazil and their counterpart in human genome.

Aline Beatriz Mello Rodrigues[1], Ana Carolina Ramos Guimarães[2]

1 INSTITUTO OSWALDO CRUZ
2 FIOCRUZ-IOC

Abstract

Leishmaniasis is a public health problem in several parts of the world, due to its wide distribution and high prevalence. Infection caused by parasites of the genus Leishmania produces in humans a set of clinical syndromes that can generate multiple or single skin ulcers, the impairment of the upper airway mucous membranes and in the viscera. In Latin America, the disease has been found in at least 12 countries, 90% of which occur in Brazil's Northeast, Central West and North regions. The treatment currently employed is extremely toxic and has a high cost to the patient, which limits its use in endemic areas. The identification of new therapeutic targets with critical importance in the survival of the parasite is aimed at the development of new drugs more effective and less aggressive for humans. In this context, the enzyme fumarate hydratase (FH) is a promising molecular target, since the parasite and Homo sapiens enzymes are analogous, i.e., descend from different ancestors and by convergent evolution have the same enzymatic function. Studies in the genus Leishmania genome point to two genes, which encode fumarate hydratase (EC 4.2.1.2). This enzyme catalyzes the reversible hydration of fumarate in S-malate, and recent studies show the importance of this enzyme for the parasite viability, which makes it a potential target for the planning of compounds with leishmanicidal action. In this perspective, this project highlights the use of computational methodologies to propose molecular models for species of Leishmania genus that can be found in Brazil. FH isoforms sequences of 3 species of Leishmania genus (L. braziliensis, L. guyanensis and L. infantum) were retrieved from TriTrypDB and UNIPROT. A FH crystallographic structure of L. major (PDB ID: 5L2R) was used as template. Subsequently, 3D models were generated using comparative modeling methods. The comparison between the sequences of the class I fumarate hydratase among some species of the genus Leishmania indicates that the residues of the catalytic site remain totally conserved, which suggests the possible inhibition of the enzyme for several species of this genus. The analysis of the structures of the parasites and host enzyme shows differences in the catalytic residues involved in the reaction. The results that will be obtained with this project may contribute to the development of new drugs against leishmaniasis.

Funding: Instituto Oswaldo Cruz

Identification of Staphylococcus aureus secretome protein signature using logistic regression to distinguish its role in interaction with the host

Ana Carolina Barbosa Caetano[1], Sandeep Tiwari[2], Núbia Seiffert[3], Vasco A de C Azevedo[4], Thiago Luiz de Paula Castro[3]

1 UFMG
2 INSTITUTE OF BIOLOGICAL SCIENCE, UFMG, BELO HORIZONTE
3 UFBA

Abstract

Staphylococcus aureus is a Gram-positive pathogen and the major causing agent of mastitis in ruminants worldwide. Mastitis is often difficult to cure, leading to important losses in farm productivity. Animal isolates of S. aureus are commonly categorized into specific clonal complexes, supported by strong genetic evidences of host-dependent specialization. The extracellular proteins produced by the pathogen are known to play a role in communication with the host and comprise the arsenal used to establish infection. In this context, variant extracellular proteins are expected to be found among different groups of S. aureus. These variants are likely to be correlated with host specificity. The variety and specificity of protein structure and function depends on sequence diversity. When combined, the twenty naturally occurring amino acid residues can form 8,000 triplets and 160,000 quadruplets. A single triplet change may affect protein folding, catalytic domains, protein-ligand interactions, and protein-protein interactions. In this study, we aimed to identify amino acid triplets found in the exoproteomes of S. aureus isolated from ewe (strain O11) and bovine (RF122). Firstly, protein sequences of both strains were downloaded from National Center for Biotechnological Information (NCBI). Then, the SurfG program was used to predict the cellular localization of proteins. 91 and 96 secreted proteins were selected for strains O11 and RF122, respectively. To analyze amino acid triplets occurrence in the two strains, the logistic regression method was applied using the MATLAB software, version R2017a. As result, we found the triplets 'DQA', 'TRI', 'PVS', 'IDV' and 'DVN' in strain RR122, whereas the triplets 'MMK', 'KMK', 'MKM', 'VQA' and 'TRV' were found in O11. Furthermore, the proteins containing most of these triplets were identified. The truncated methicillin resistance-related surface protein (CAI81729.1) and the hypothetical protein tagged as EGA96785.1 were found in the RF122 and O11, respectively. The outcome of this work could facilitate in-silico functional characterization and the study of the differential interaction of the two strains with their respective hosts. We plan to include more strains from different S. aureus groups and further characterize the interaction with different hosts.

Funding: CAPES, TWAS, CNPQ, FAPEMIG

A Parallel Bioinspired approach to the Protein Folding Problem using a coarse-grained model

Andrey Cabral Meira[1], César Manuel Vargas Benítez[1]

1 UTFPR

Abstract

One of the great challenges of Bioinformatics and Molecular Biology is to predict the structures of proteins from their amino acid chain and their chemical interactions. The Protein Folding Problem (PFP) has been constantly studied aiming to find efficient solutions. The importance of studying the folding of proteins may be justified due to the fact that, according to the contemporary knowledge in Biology, errors at some point of the folding might cause diseases such as Cancer, Alzheimer's, Cystic Fibrosis, bovine spongiform encephalopathy (mad cow disease) among other diseases. The present project aims an analysis of the PFP based on bioinspired approaches (such as Genetic Algorithms, Artificial Bee Colony, Differential Evolution) and Heuristics with the use of Parallel Computing with GPUs (Graphics Processing Unit), which are used for the processing of the calculations and folding dynamics. The specific goal of the project is to predict the functional structure of a protein from its primary structure and its hydrophobic and hydrophilic interactions. The use of GPUs become interesting due to the fact that with the parallelization of the calculations it is possible to divide the problem in several fractions to improve the processing time. The Coarse-Grained model, which is an alternative to the atomic model, consists on the representation of amino acids as particles with interaction sites and will be used aiming the reduction of the computational effort. It is known that the PFP is defined as an NP-Difficult problem, which also justifies the use of heuristics and Parallel Computing in the methodology. As result of the study, it is expected to produce an approach that can efficiently contribute to PFP studies with satisfactory computation for replication.

Funding: PPGBIOINFO, CNPq

Integrated model of the mRNA translation and the amino acid chain folding within the ribosome tunnel

Bárbara Zanandreiz de Siqueira Mattos[1], A.p.f. Atman[2], Anton Semenchenko[1]

1 CENTRO UNIVERSITÁRIO NEWTON PAIVA
2 CEFETMG

Abstract

The ribosome is the main facilitator of the protein synthesis process. It influences the formation of the secondary structure of the nascent polypeptide chain within its exit tunnel. The ribosome exit tunnel is an active factor in the formation of the amino acid chain secondary structures due to the various mechanisms of interaction between the nascent chain and the elements that form the walls of the tunnel. In addition to the ribosome influence onto the nascent chain, there is a negative feedback from the co-translational folding process within the tunnel and the mRNA translation by the ribosome. The presented project intends to construct the mathematical representation of the interaction between the amino acids and the ribosome exit tunnel and integrate the co-translational folding model with the real-time mRNA translation by the ribosome. The project evaluates the hypothesis that the spatial limitation of the tunnel geometry and, specifically, the constriction location within the tunnel, interfere on the nascent chain secondary structure. As the result, we demonstrate the improved the ability to predict the distribution of the amino acid structures within the ribosome exit tunnel using only computational tools calibrated by the cell-free protein synthesis process. This result is achieved by the integration of the real-time mRNA translation and the simulation of the polypeptide chain folding within the ribosome exit tunnel. Specifically, we visualize the secondary structures of the nascent amino acid chain in the process of formation within the ribosome exit tunnel. The main technique to perform this work is the agent-based modeling with the support of the NetLogo© 3D computational tool. The proposed model represents the process of amino acid chain folding by the attachment of one amino acid at a time and the propagation of the nascent chain inside the exit tunnel of the ribosome. The timing of the amino acid insertion into ribosome exit tunnel is determined by the real-time mRNA translation by the ribosome and validated by the experimental data. The three-dimensional positioning of the nascent chain within the tunnel is represented as a stochastic process of the orientation changes of the individual elements, according to the degree of freedom allowed by the chemical bonds within the amino acids. The simulated results are compared and validated by the existing crystallography experiments.

Funding: None

In-silico Structural Characterization of Variants Found in PCSK9 gene Identified in Familial Hypercholesterolemic Patients

Bruna Los[1], Jéssica Bassani Borges[2], Gisele Medeiros Bastos[3], André Arpad Faludi[3], Rosário Dominguez Crespo Hirata[1], Mario Hiroyuki Hirata[2]

1 FACULTY OF PHARMACEUTICAL SCIENCES - USP
2 FACULTY OF PHARMACEUTICAL SCIENCES - USP AND DANTE PAZZANESE INSTITUTE OF CARDIOLOGY
3 DANTE PAZZANESE INSTITUTE OF CARDIOLOGY

Abstract

Familial Hypercholesterolemia (FH) is a genetic disorder of lipoprotein metabolism, mainly caused by mutations in three genes, LDLR, APOB, and PCSK9. PCSK9 acts regulating low density lipoprotein (LDL) levels by binding to LDL receptor (LDLR) and escorting it towards intracellular degradation compartments. Gain-of-function mutations in PCSK9 increase its proteolytic activity, reducing LDLR concentration, therefore resulting in high levels of LDL cholesterol in the plasma. Loss-of-function mutations lead to a higher concentration of the LDLR, resulting in lower LDL cholesterol levels. The aim of the present project is an in silico and in vitro characterization of the effect of variants in PCSK9 gene identified in FH patients. Forty-eight FH patients were sequenced using Next Generation Sequencing. The data were aligned to the reference genome using Burrows-Wheeler Aligner (BWA) and variant calling was performed using Genome Analysis Toolkit (GATK). After this, nine missense variants were identified in PCSK9 gene. Between them, four were chosen to further analysis because were visible in the crystal structure and presented MAF below 5% in three databases. Crystal structures of wild type PCSK9 and LDLR were retrieved from Protein Data Bank (PDB code: 2P4E and 1N7D, respectively) and site-directed mutagenesis was performed using PyMOL v. 1.8.6.2. to generate the following PCSK9 variants: R237W, A443T, R469W and Q619P. Structural analysis of molecular interactions of PCSK9 and its variants with LDLR was performed by protein-protein docking via ClusPro. The PCSK9-LDLR complexes were visualized using PyMOL v. 1.8.6.2. For R237W and R469W it was observed a possible conformational change that could increase the affinity of PCSK9 for LDLR, when compared with the wild type. In both cases, the arginine to tryptophan change allowed an interaction with a LDLR region featured by a hydrophobic pocket. For A443T and Q619P no conformational changes were observed, and both variants showed only interactions with PCSK9 amino acids itself, suggesting theses variants are probably neutral. R237W was already defined as a loss-of-function mutation by in vitro studies; however, no functional assays were performed on R469W. As previous genetic association studies indicate that R469W is a gain-of-function mutation, and led by our in silico result, an in vitro characterization will be conducted to further understand the possible pathogenicity of the R469W.

Funding: CNPq

Aspergillus fumigatus : computational characterization of UBP14 deubiquitinase

Carlos Bruno de Araujo[1], Juliana da Silva Viana[1], Natália Silva da Trindade[1], Polyane Vieira Macêdo[1], Matheus de Souza Gomes[1], Enyara Rezende Morais[1]

1 UFU

Abstract

Invasive Pulmonary Aspergillosis (IPA) is a disease that has a high mortality rate ranging from 30 to 90%, caused by the opportunistic fungus Aspergillus fumigatus. This pathogen is highly virulent due to the fact that it has some mechanisms of resistance to adverse situations. Eukaryotes have an important non-functional protein degradation system, which is known as Ubiquitin-Protoeasome System (UPS). This system is composed of several enzymes, the deubiquitinases (DUBs) play a fundamental role during the process because they act in the breakdown of the bonds between ubiquitin molecules. UBP14 participates by releasing ubiquitin from polyubiquitin chains which are not anchored to the substrate, such action re-establishing free Ub levels in the cell medium and preventing inhibition of the proteasome by binding of such chains. The UPS is very important for the survival of several organisms besides it has not yet been characterized in A.fumigatus, so the objective of this work was to characterize the UBP14 deubiquitinase in this pathogen. Initially the Saccharomyces cerevisiae UBP14 protein sequence was used to identify the corresponding ortholog in A. fumigatus from ASPGD (aspergillusgenome) database, then using BLAST it was possible to identify the best orthologs for the prediction of the conserved domain in the Pfam, the amino acid residues from the catalytic site in CDD, for multiple alignment in CLUSTALX and phylogenetic analysis in MEGA5.2. Afu2g06330 was identified as UBP14 in A. fumigatus, which has 783 amino acid residues, the UCH domain and the Peptidase C19 catalytic site. The results obtained in this study demonstrated the presence of the studied protein in A. fumigatus. These characterizations enable the use of this protein as a potential molecular target.

Funding: FAPEMIG, CNPq, UFU and CAPES

IN SILICO MODELING OF THE C2H2 ZINC-FINGER DOMAIN OF THE GLI3 TRANSCRIPTION FACTOR

Cinthia Caroline Alves[1], Eduardo Antônio Donadi[1], Silvana Giuliatti[1]

1 RIBEIRÃO PRETO MEDICAL SCHOOL, USP

Abstract

The intracytoplasmic glioma-associated oncogene-3, GLI-3, is a protein that belongs to the zinc-finger protein family and has dual function as a transcriptional activator and a repressor of the Sonic Hedgehog pathway. The full-length GLI3 form (GLI3-190kDa) after phosphorylation and nuclear translocation, acts as an activator (GLI3A), while GLI3R (GLI3-82kDa), its C-terminally truncated form, acts as a repressor. Since both protein forms present an important C2H2 type zinc-finger domain that allow protein binding at the DNA sequence, it is necessary to generate the complete 3D structure of DNA-binding-domain of the GLI3 to better understand its role as a transcription factor. The C2H2 domain of the GLI3 protein compasses the residues 480-632 (UniProt code: P10071 - http://uniprot.org/) and this sequence was used to further analysis. Its secondary structure was predicted using the PSIPRED webserver (http://bioinf.cs.ucl.ac.uk/psipred). The tertiary structure was modeled by homology using the Modeller 9.19 software, and the 2.6 crystal structure (PDB code: 2GLI) available on PDB (http://rcsb.org/pdb) was the chosen template to generate 5 models. Quality assessment of these models was performed by torsion angles analysis (using PROCHECK and PDBSUM), visual analysis and distance evaluating (using CHIMERA software). Energy minimization and equilibrium steps (5 ns) of the chosen best model was performed using the GROMACS 4.6.5 software, which was also used to perform the molecular dynamic simulation. Homology modeling allowed satisfactory GLI3 C2H2 domain prediction models which presented good quality torsion angle assessment. The chosen best model presented 89.9% of the residues in the core region of phi-psi torsion angles, while 8.4% of the residues are in allowed regions, and it showed the lowest root-mean-square deviation (RMSD) of 0.729 after model-template alignment. After energy minimization and equilibrium of the chosen model, the molecular dynamic simulation was done. In conclusion, the initial quality assessment showed a satisfactory 3D structure generated of the GLI3 C2H2 domain by homology modeling, which can be used as a template for modeling DNA binding domains and to perform protein-DNA interaction studies in the future.

Funding: CAPES, CNPq

A new method based on structural signatures to propose mutations for enzymes β-glucosidase used in biofuel production

Diego Mariano[1], Raquel Melo Minardi[1]

1 UFMG

Abstract

β-glucosidases (E.C. 3.2.1.21) are key enzymes in the second-generation biofuel production process. They act synergically with endoglucanases and exoglucanases to convert cellulose of biomass in fermentable glucose used in biofuel production. However, it has been reported in the literature that the majority of known β-glucosidases is inhibited by high concentrations of glucose. Hence, it has increased the search for mutations that improve the activity and glucose tolerance. In this study, we present a method to propose mutations for enzymes β-glucosidase that may improve the activity and tolerance to glucose inhibition. Our method is based on structural signatures: numerical representations of proteins extracted from the number of pairwise residues. We hypothesized that proteins with similar structural signatures of catalytic pockets present similar characteristics. Hence, mutations that approximate non-tolerant β-glucosidase structural signatures of other enzymes classified in the literature as glucose-tolerant may improve the activity of these enzymes. We used Euclidian distance to calculate signature variations. If the signature variation was negative, the distance between signatures reduced, so we consider as a beneficial mutation. If the signature variation was positive, the distance between signatures increased, so we consider as a not beneficial mutation. We collected 27 mutations in β-glucosidases from literature and classified them in beneficial or not beneficial based on the experimental effects reported. Then, we calculated the signature variation for every mutation and compared the predicted result with the real result. We obtained a precision value of 0.89. In addition, we proposed 15 mutations for Bgl1B, a non-tolerant β-glucosidase extracted from marine metagenome. We detected experimental data in the literature for three of these mutations: H228C, H228T e H228V. The experimental data demonstrate that these mutations improve the activity even in high glucose concentrations. These results show that our method is efficient to detect mutations that increase the activity of β-glucosidases and it can help to produce new mutant enzymes that may improve the second-generation biofuel production.

Funding: CAPES

Evaluation of the molecular impact of an exclusive aminoacid substitution of Saccharomyces cerevisae more tolerant to ethanol strains: a molecular dynamics approach.

Guilherme Ferreira Luz[1], Guilherme Targino Valente[1], Rafael P. Simões[1]

1 UNESP - STATE USP

Abstract

The society claim for a new way of acquiring energy. Since the world demands an independence of fossil fuels, ethanol is raising as a great alternative to the global issue. The yeast Saccharomyces cerevisae is the microorganism most used for ethanol production because of its great fermenting capacity and also a great resilience in this process. Although, the ethanol concentration on the production is one of the most limiting factors of the industry, once the product is toxic for living cells. Saccharomyces cerevisae has a plenty of different strains, which could explain why some strains are more tolerant to ethanol than others. The study of mutations in different strains could give us a path to understand the ethanol tolerance phenotype. In this context, the current project aims to understand the phenomenon of ethanol tolerance selecting a particular protein applying bioinformatic tools and analyzing it by molecular modeling approach. The candidate was chosen using alignment analysis of the whole proteome of five different strains (S288C, BY4741, BY4742, SEY6210 and X2180-1A) being the X2180-1A, BY4741 and BY4742 the most tolerant ones. Thus, the ADH1 (alcohol dehydrogenase) protein is a protein with an activity of alcohol catalysis, breaking the alcohol and turning it to acetone or aldehyde groups. The ADH1 was chosen since it has a single mutation exclusive of the most tolerant strains. That mutation changes a Glutamine for a Glutamate, adding an extra electron to the chain. The ADH1 protein is a dimeric protein with 2 zinc ions attached to the chain (1 catalytic zinc and 1 structural) and 1 NAD+ (a coenzyme factor). The molecular analysis showed the observed mutation changed the electronic structure of the molecule catalytic site, which could improving the catalytic zinc activity by increasing its oxidation potential. The molecular dynamics simulations using the Charmm force field showed that the mutated structures have similar conformation energy as the non-mutated ones, and it can be considered stable molecular structures. The electronic distribution has been performed using the Gaussian09 package and the oxidation potential of the catalytic site has been calculated as well. If the hypothesis of faster oxidation to be confirmed, we could be able to prove that the single mutation in ADH1 protein might be responsible for ethanol tolerance increasing. The results could provide knowledge for new target genes for genetic engineering of Saccharomyces cerevisae to increase the ethanol tolerance.

Funding: UNESP - State University of Sao Paulo (Botucatu)

Virtual Screening of potential inhibitors for the Alpha-Amylase and Alpha-Glycosidase by shape based model and docking

Heitor Cappato[1], Nilson Nicolau Junior[1], Foued Salmen Espindola[1]

1 UFU

Abstract

Natural antioxidants compounds have been associated with reduction of postprandial hyperglycemia by blocking enzymes involved in the carbohydrates digestion, such as alpha-amylase and alpha-glycosidase. Furthermore, preventing or delaying the absorption of glucose by inhibiting glycoside hydrolases in the digestive organs may represent a promising approach in the treatment of diabetes and its complications. Thus, the aim of this work was search for new natural compounds with pharmacological potential to inhibit this glycoside hydrolases based on the shape and color based model and docking. The shape and color modeling was performed with the aid of vROCS 3.2.0.4, this model contains information about shape and chemical properties extracted from the acarbose molecule. The ligand library used in this research are originated from ZINC database, that have been carefully selected a natural compounds subset, totaling 180.303 compounds. In order to perform the virtual screening, the ligand library was prepared with the OMEGA 2.5.1.4, which was used to generate conformer libraries. Pharmacophore model validation and virtual screening of the conformer libraries were performing using vROCS. The pharmacophore model was previously validated using the ROC (receiver operating characteristic) curve and AUC (area under the curve). In order to generate the ROC curve and the AUC value, biologically active ligands against alpha-amylase (PDB id: 1SMD) and alpha-glycosidase (PDB id: 1OBB) were obtained from ZINC database, and the decoys were generated on the DUD-E online platform. After validation, the conformer library previously generated was submitted to the shape and color model and the top 500 ligands of each, based on the TanimotoCombo score, were selected. The best-scored ligands were used to perform a molecular docking against human alpha-amylase and alpha-glycosidase using the autodock vina 1.1.2, generating three potential inhibitors that are of different class of compounds usual inhibitors.

Funding: FAPEMIG, CNPq, UFU and CAPES

Detection and prediction of premature stop codon using mass spectrometry data at the protein level

Karla Cristina Tabosa Machado[1], Andre Fonseca[1], Sandro Jose de Souza[1], Gustavo Antônio de Souza[1]

1 UFRN

Abstract

The volume of public data regarding proteomics has increased significantly in the last few years, allowing to use its potential to improve the annotation of genomics data. The key stage in proteomics is to identify peptides and proteins, initially through a comparison of collected mass spectrometry (MS) data and theoretical sequences in a database. However, protein sequence databases report mostly a no-redundant number of known isoforms, while individual polymorphic variations are not represented. Nonsense mutations are characterized by the premature appearance of the stop codon in a gene, which could produce defective proteins. The objective of this paper was to define a computational approach which could allow the prediction, of such mutations in proteomic datasets, without previous knowledge of the genome of the sample and consequently, only using a reference protein sequence database for protein identification. It was proposed that when nonsense mutations are present, one could track the quantitative profile of each peptide of a given protein, in order to detect a drop of sequence coverage at the protein c-terminal, that could be explained by the presence of a premature stop codon. This method was developed with the use of Perl language programming scripts which divide the identified proteins with sequence coverage above 30% into bins, in which each bin is 5% of the size of the subject protein. The script verifies peptides structures to ensure that they are totally or partially included inside of each bin and then adds all spectral counts associated to each peptide. Public MS data from three publications investigating the proteomes of cell lines, ovary cancer and colon cancer were re-analyzed using the approach developed here. Among the findings, genes such as BUB3, CALR and PRMT5 appeared mutated in certain samples of patients with cancer. This data will be validated at a later date to confirm the presence of the stop codon in the gene.

Funding: UFRN

Spatial representation of amino acid composition divergence in homologous protein families

Lucas Carrijo de Oliveira[1], Néli José da Fonseca Júnior[1], Lucas Bleicher[1]

1 UFMG

Abstract

Homologous protein families can be assessed by multiple sequence alignments (MSA), wherein each column represents an evolutionarily corresponding position among homologous proteins. Conserved positions, meaning invariable sites, indicate evolutionary constraints in amino acid substitutions, generaly due to structural and/or functional importance of such positions. Besides the fully conserved ones, there are some positions that are specificaly conserved in functional subclasses eventualy present in a family. In the same way, as one moves toward a phylogenetic tree, from root to leaves, some residues appears as being specificaly conserved in each clade, while others remain variable or unspecificaly conserved. By representing each residue (here calling residue a given amino acid in a specific position, like "His37") by the set of all sequences in a MSA having such a residue, and calculating the conditional probabilities of finding all other ones given the presence of that residue (e.g., probability of finding Asp71 in sequences having His37), it is possible to compare all possible sets of sequences on the basis of their amino acids composition. The present work introduces a distance based method to represent, in the N-dimensional space, the evolutionary divergence of amino acid composition in homologous protein families. For each residue, the method takes a specific sub-alignment (e.g, the subset of sequences in a MSA having that residue) and considers each column as a 20-dimensional vector, being each dimension the conditional probability of finding, at that position, each of the 20 amino acids. This way, each sub-alignment is represented as a set of 20-dimensional points in space. Two sub-alignments can than be compared by calculating the root mean square deviation (RMSD) between these two sets of points. An all against all distance matrix is generated and, by singular value decomposition (SVD), it is possible to define N-dimensional spatial coordinates from this distance matrix. By ploting these coordinates in a tridimensional Cartesian plane, one can visualize the pattern of amino acid composition divergence in homologous protein families, from more conserved to more specific residues, passing toward variable or unspecificaly conserved ones. Colouring points by frequency in MSA of their respective residues helps visualization of such an effect.

Funding: CAPES, CNPq

Structural features of HIV-1 Integrase mutations in patients and in vitro samples treated with strand transfer Inhibitors

Lucas de Almeida Machado[1], Ana Carolina Ramos Guimarães[1]

1 FIOCRUZ

Abstract

Acquired immunodeficiency syndrome (AIDS) is one of the greatest health challenges in modern medicine. According to the UNAIDS, in 2014 nearly 35 million people were living infected with the HIV (Human immunodeficiency virus) worldwide, of which 734 thousand live in Brazil - where HIV-1 is the predominant type . In spite of the reduction of AIDS mortality due to the relative success of HAART (highly active antiretroviral therapy), many patients do not respond to the treatment with protease and reverse transcriptase inhibitors, and the HIV-1 integrase inhibitors are part of the last resources in therapy. HIV-1 integrase is a 288 residue enzyme responsible for the integration of the viral DNA into the host genome. In the last years the integrase inhibitors Raltegravir and Elvitegravir were widely used in therapy, however, due to the high rates of resistance mutations against these inhibitors, the second generation inhibitor Dolutegravir was implemented. In spite of the fact that Dolutegravir has higher genetic barriers to resistance, many Dolutegravir resistance mutations have been described recently. In the present work, we attempted to investigate structural features of the mutations present in treated individuals and check whether or not such mutations were already described in the literature and also analyze structural features of the positions mutated. Our databank of HIV-1 integrase sequences was built of patient samples from the HIV drug resistance database and in vitro samples. The databank was separated into groups based on the inhibitors each patient or sample received. For each group, the frequency of missense mutations at each position was calculated. To evaluate the structural features of each highly mutated residue, a comparative model was built with Modeller 9.17, using as templates a structure of the integrase tetramer in complex with DNA (5u1c) and a two-domain structure of the integrase (1e4x), the model with the lowest dope score was refined and validated. Many of the highly mutated sites were not cited in the literature as involved in resistance or accessory mutations, and many of the positions described as involved in resistance do not feature the top mutated sites. At least 16 mutations not described in the literature appear close to protein-DNA interface, to the active site or to residues that play key roles in DNA anchoring. Our data suggest that residues not described before may play a role in resistance, however further studies are needed to determine if such positions are important for viral fitness.

Funding: CAPES

Identification and computational evaluation of possible allosteric and competitive inhibitors of human PEPCK-M: an alternative therapy for lung carcinoma

Luiz Phillippe Ribeiro Baptisa[1], Vanessa de Vasconcelos Sinatti Castilho[1], Ana Carolina Ramos Guimarães[1]

1 FIOCRUZ-IOC

Abstract

Cancer is the second largest cause of death in the world, posing a huge problem for modern medicine. The increase in the number of cells, the result of uncontrolled cell division, leads to an increasing requirement of glucose consumption. Tumor growth under conditions of metabolic limitation, especially with decreased glucose availability, is common, suggesting that tumor cells exhibit high metabolic plasticity. Central to this adaptation, there is the enzyme phosphoenolpyruvate carboxykinase (PEPCK) that participates in the initial phase of gluconeogenesis. This enzyme acts on the reversible formation of phosphoenolpyruvate (PEP) from oxaloacetate (OAA). Due to its gluconeogenic function, the differential expression of cytoplasmic (PEPCK-C) and mitochondrial isoforms (PEPCK-M) is independently associated with different types of cancer. Recent studies have shown that this change is critical in lung cancer, in which tumor cells submitted to low glucose levels increase the expression of PEPCK-M. Also, inhibition or the knockdown of the PEPCK-M enzyme in lung tumor cell cultures led to increased cell death and apoptosis. These studies show the importance of this enzyme to the prevalence of cancer. In this work, we intend to explore the enzyme PEPCK-M as a target for a computer-aided drug design that can be used in the therapy of lung cancer. Although very important, there are no 3D structures deposited in the PDB for PEPCK-M (unlike its isoform, PEPCK-C). For this reason, we performed a comparative modeling using the program MODELLER. Comparative analyses between the two isoforms have shown that they are very similar - presenting high conservation between active and allosteric site residues. The electrostatic potential analysis, performed with APBS, also indicated strong similarities - both enzymes have an overall electropositive active-site cleft. We also conducted redocking experiments with Glide XP and Autodock Vina to test which program is the most suitable for future experiments. The observation of mutations commonly found in lung cancers was another important analysis. This experiment used the mutation database COSMIC. The most frequent mutations were mapped on the PEPCK-M structure. These results suggest that the PEPCK-M enzyme has structural characteristics like those shown in PEPCK-C - validating the use of inhibitors described for the cytoplasmic enzyme. Also, the localization of common mutations in lung cancer, both in the catalytic cleft and allosteric site, favor the search for specific inhibitors for the mutated PEPCK-M.

Funding: CNPq, Fiocruz

In silico improvement of the cyanobacterial lectin microvirin and Mana(1-2)Man interaction

Adonis Lima[1], Andrei Santos Siqueira[1], Luiza Möller[2], Rafael Souza[3], Alex Ranieri Jerônimo Lima[1], Ronaldo Correia da Silva[1], Délia Cristina Figueira Aguiar[1], João Lídio da Silva Gonçalves Vianez Junior[3], Evonnildo Costa Gonçalves[1]

1 UNIVERSIDADE FEDERAL DO PARÁ
2 FACULDADE INTEGRADA BRASIL AMAZÔNIA
3 INSTITUTO EVANDRO CHAGAS

Abstract

Given the impact of human immunodeficiency virus (HIV) infection, a portion of the scientific community has been dedicated to the development of drugs capable of preventing the virus from entering the host cell, thus preventing the infection of new individuals. This study aims to perform in vitro microvillin analysis (MVN), a lectin produced by the Microcystis aeruginosa cyanobacterium, aiming at optimizing its binding affinity for the viral gp120 protein, the protein that mediates virus entry into CD4 + T cells. The nucleotide sequence of this work was obtained from a genomic analysis of the Microcystis aeruginosa CACIAM 03, isolated from a surface water sample of the reservoir of the Tucuruí plant. The model was constructed by comparative modeling through the Modeller 9.16 program, having as template the Microcystis aeruginosa MVN of 2YHH PDB code (108 aa). Molecular docking of the MVN with its ligand was performed through Molegro Virtual Docker (version 5.5). The validation of the modeled target was done by analyzing the stereochemical quality, the free energy of the system and the mapping of the molecular electrostatic potential. In addition, three molecular dynamics (DM) of 210 ns were prepared using Amber16 for the refinement of the target. The alanine scanning webserver tool was used to study the importance of protein residues with its ligand. Several information acquired through these computational simulations were used to obtain a mutant. As results, the constructed target of MVN_CACIAM 03 showed 95% sequential identity as compared to the 2YHH template. In the generated MVN_CACIAM 03, 97% of the residues were found in favorable regions, according to the Ramachandran graph. Molecular mooring decreased the energetic state of the complex, which also confirmed the interactions described in the literature. The RMSD values ??of the mannose and the interaction site were very stable during the three trajectories of 210 ns. Calculations of the occupation time of the hydrogen bonds were made for the residues that showed interaction in the MVN_CACIAM03 complex and mannose. And the generated mutant (Thr82Arg), after computational studies, showed to be more efficient during the process of receptor-ligand interaction.

Funding: CAPES

Low Molecular Weight Phophatases: Coevolved residues and a Mutation Database

Marcelo Querino Lima Afonso[1], Néli José da Fonseca Júnior[2], Lucas Bleicher[3]

1 UFMG

Abstract

The Low Molecular Weight Phosphatase fold protein family has importance in various eukaryotic and prokaryotic cellular signalling networks. Its proteins influence various diseases such as cancer, diabetes and tuberculosis. Many enzymatic functions are displayed by this protein family, the most important being the dephosphorylation of Tyrosines, Arginines, Ribulosamines and Erythrulosamines, and the reduction of the Arsenate ion to Arsenite. Multiple characterizations, structures and site-directed mutagenesis related to this protein family exist up to this date. In this work, we aimed to describe the biological functions related to the correlated and anticorrelated residue sets extracted from the Pfam database Multiple Sequence Alignment for this family. We found a clear pattern of residues related to the Tyrosine Phosphatase and Arsenate Reductase classes in the network and describe here new possible important positions to be explored in future experimental studies.

Funding: FAPEMIG, CAPES

Identifying specificity determinant residues through decomposition of protein families affiliation network

Néli José da Fonseca Júnior[1], Lucas Carrijo de Oliveira[1], Marcelo Querino Lima Afonso[2], Lucas Bleicher[1]

1 UFMG

Abstract

Affiliation networks are widely used in the context of social and ecological systems. In the present work, we embrace the state of art in this field in order to apply it in the mapping of amino acid coevolution patterns. The goal of this project consists in, given a multiple sequence alignment, predict patterns of local residue conservation that may be related to some specificity (functional, structural or taxonomic). A bipartite network is modeled from a multiple sequence alignment, in a way that each protein is connected with their respective residues. This network is then projected to a residue monopartite representation and its backbone is extracted in order to remove statistically insignificant edges. Finally, the resulting network is decomposed into communities of residues that are more likely to co-occur. We evaluated seven methods for network sparsification with simulated data. These virtual alignments were randomly generated with functional and secondary evolutionary constraints. Experiments with real data were also performed using the HIUase/Transthyretin family and the G protein-coupled receptor, rhodopsin-like family. The results showed that most of the sparsification methods evaluated could in fact rise coevolution patterns in this type of networks. We detected specificity determinant residues for both subclass of the HIUase/Transthyretin family using either filters or weighting to treat the alignment bias. Several functional subclasses were also identified in the GPCR analysis. The methodology presented here is fast and useful to analyze specificity determinant sites, functional subclasses and local conservation residues. This pipeline can be used with large multiple sequence alignments as the obtained from Pfam. Depending on the method used to extract the backbone, anti-correlations could also be observed. Stereochemical correlations can also be identified by generating multiple networks with different amino acid alphabets.

Funding: Capes

Functional prediction of stress-modulated proteins of Deinococcus radiodurans

Ricardo Valle Ladewig Zappala[1], Manuela Leal da Silva[2], Pedro Geraldo Pascutti[1], Claudia de Alencar Santos Lage[1]

1 UNIVERSIDADE FEDERAL DO RIO DE JANEIRO
2 INSTITUTO NACIONAL DE METROLOGIA, QUALIDADE E TECNOLOGIA

Abstract

The Deinococcaceae group comprises some of the robust known extremophilic bacteria. Attempts have specially focused on responses against extreme doses of gamma radiation or desiccation to explain survival of Deinococcus radiodurans against them. Many defensive mechanisms were shown to exist in D. radiodurans, and transcriptomes already performed in response to gamma radiation and desiccation revealed that some genes were transcribed to proteins of undefined functions, while others have never been expressed under those conditions. Therefore, it is expected that such genes with unknown functions could code for novel resistance proteins to those extreme conditions. The present study aims to identify and perform function prediction for hypothetical, unique proteins of D. radiodurans, without similarity to any other known protein. Sequences of a group of 26 proteins, with 23 expressed in D. radiodurans after radiation or desiccation stresses were retrieved, which hypothetical functions were predicted by the best scores after BLAST alignments and CD-search. Information about the proteins was gathered through alignments against Uniprot and PDB databases. Using molecular modeling tools as I-TASSER, SWISS MODEL and MODELLER, 3D models were built for all hypothetic proteins and they were mainly evaluated by Ramachandran's Plot, RMSD and DOPE score. The best models were then submitted to structural classification on SCOP and CATH servers. This approach enabled us to better infer about the function of twenty candidates for new extremophilic proteins, of which thirteen went through comparative modeling with multiple templates. Three seemed to belong to the group of intrinsically disordered proteins, and three have not aligned to any proper templates by comparative modeling. Among the seven predicted proteins using structures from SWISS MODEL are: DR0438, a DNA binding protein; DR1263, a N-glycosidase; DR1314, a photosystem-like transmembrane protein; DR1370, a structural lipoprotein; DR2073, a kinase; and DR2441, an acetyl-transferase. Among the 26 analyzed proteins, the most interesting one appears to be the DR0491 gene product, showing 25% identity, 41% similarity, and covering 90% of the sequence correspondent to the Escherichia coli heat shock protein Hsp31. This may represent an essential role on catalysis of damaged proteins, as well as proper folding assistance on other unstable proteins. After several steps of investigation, modeling and structural analyses, complementary tools such as phylogeny, molecular dynamics and molecular docking were also performed to strengthen the significance of the observed results, and this particular resistance toolbox with novel and exclusive proteins was referred as the "Black Box Genome of D. radiodurans".

Funding: CNPq

Functional analysis of hypothetical proteins unveils putative metabolic pathways, essential genes and Therapeutic drug and vaccine target in Trypanosma cruzi: A Bioinformatics Based Approach

Rodrigo Profeta Silveira Santos[1], Priya Trivedi[2], Neha Jain[2], Sandeep Tiwari[3], Syed Babar Jamal Bacha[4], Arun Kumar Jaiswal[5], Thiago Luiz de Paula Castro[6], Núbia Seiffert[6], Siomar de Castro Soares[7], Artur Silva[8], Vasco A de C Azevedo[1]

1 UFMG
2 DEVI AHILYA UNIVERSITY
3 INSTITUTE OF BIOLOGICAL SCIENCE, UFMG
4 1. INSTITUTE OF BIOLOGICAL SCIENCE, UFMG
5 INSTITUTE OF BIOLOGICAL SCIENCE, UFMG; DEPARTMENT OF IMMUNOLOGY, MICROBIOLOGY AND PARASITOLOGY, INSTITUTE OF BIOLOGICAL SCIENCES AND NATURAL SCIENCES, UFTM
6 UFBA
7 DEPARTMENT OF IMMUNOLOGY, MICROBIOLOGY AND PARASITOLOGY, INSTITUTE OF BIOLOGICAL SCIENCES AND NATURAL SCIENCES, UFTM
8 UFPA

Abstract

The protozoan Trypanosoma cruzi is the etiological agent of Chagas disease, a major chronic, systemic, parasitic infection. The disease affects about 8 million people in Latin America, of whom 30-40% either has or will develop cardiomyopathy, digestive mega syndromes, or both. Currently, there are neither effective drugs nor vaccines for the treatment or prevention of the disease. The current synthetic drugs such as nifurtimox (a nitrofuran derivative) and benznidazole (a nitroimidazole derivative), are associated to severe side effects, including cardiac and/or renal toxicity and as well not effective, which accounts for the need to search new effective chemotherapeutic and chemo prophylactic agents against T. cruzi . Therefore, due to their low efficacies and the resistance developed by the bug to these medications, there is an urgent need to consider newer species-specific targets. Approximately 50% of the predicted protein-coding genes of the Trypanosoma cruzi CL Brener strain are annotated as hypothetical or conserved hypothetical proteins.Here in this work, we have attempted to assign probable functions to these hypothetical sequences present in this parasite, to explore their plausible roles as druggable receptors. Thus, putative functions have been defined to 491 hypothetical proteins, which exhibited a GO term correlation and PFAM domain coverage of more than 50% over the query sequence length. We tried to find out if our 491 sequences were showing any similarity

Proteome scale comparative modeling for conserved drug and vaccine targets identification in Salmonella serovers

Syed Babar Jamal Bacha[1], Jyoti Yadav[2], Neha Jain[3], Sandeep Tiwari[1], Arun Kumar Jaiswal[4], Thiago Luiz de Paula Castro[5], Núbia Seiffert[5], Siomar de Castro Soares[6], Artur Silva[7], Vasco A de C Azevedo[8]

1 INSTITUTE OF BIOLOGICAL SCIENCE, UFMG
2 SCHOOL OF BIOTECHNOLOGY, DEVI AHILYA UNIVERSITY, INDIA
3 DEVI AHILYA UNIVERSITY
4 INSTITUTE OF BIOLOGICAL SCIENCE, UFMG, DEPARTMENT OF IMMUNOLOGY, MICROBIOLOGY AND PARASITOLOGY, INSTITUTE OF BIOLOGICAL SCIENCES AND NATURAL SCIENCES, UFTM
5 UFBA
6 DEPARTMENT OF IMMUNOLOGY, MICROBIOLOGY AND PARASITOLOGY, INSTITUTE OF BIOLOGICAL SCIENCES AND NATURAL SCIENCES, UFTM
7 UFPA
8 UFMG

Abstract

Despite extensive surveillance, foodborne Salmonella enterica infections continue to cause a significant burden on public health systems worldwide. Salmonella is a food-borne pathogen that leads to substantial illness worldwide. The clinical syndromes associated with Salmonella infection are enteric (typhoid) fever and gastroenteritis, in healthy humans. Typhoid fever is caused by host-adapted S. Typhi and S. Paratyphi. Gastroenteritis is caused by serovars usually referred to as non-typhoidal Salmonellae (NTS). In this work, we used a Modelome approach for the proteome of Salmonella Typhi species. This served to bridge the gap between raw genomic information and the identification of good therapeutic targets based on the three-dimensional structures. The novelty of this strategy relies in using the structural information from high-throughput comparative modeling for large-scale proteomics data for inhibitor identification, potentially leading to the discovery of compounds able to prevent bacterial growth. The proteomes of 3 Salmonella typhimurium strains were modeled (pan-modelome) using the MHOLline workflow. Intra-species conserved proteome (core-modelome) with adequate 3D models was further filtered for their essential nature for the bacteria, using the database of essential genes (DEG). This led to the identification of essential bacterial proteins without homologs in the host proteomes. Furthermore, we investigated a set of essential host homologs proteins. We observed residues of the predicted bacterial protein cavities that are completely different from the ones found in the homologous domains, and therefore could be specifically targeted. By applying this computational strategy, we provide a final list of predicted putative targets in Salmonella typhimurium which were common to all the three serovars. They could provide an insight into designing of peptide vaccines, and identification of lead, natural and drug-like compounds that bind to these proteins. We propose that some of these proteins can be selectively targeted using structure-based drug design approaches (SBDD). Our results facilitate

In silico screening of volatile compounds which can complex with the AeagOBP1 odor-binding protein of Aedes aegypti L.

Tarcisio Silva Melo[1], Liliane Pereira de Araújo[1], Rosangela Santos Pereira[1], Thaís Almeida de Menezes[2], Wagner Rodrigues de Assis Soares[1], Bruno Silva Andrade[1]

1 UNIVERSIDADE ESTADUAL DO SUDOESTE DA BAHIA
2 UNIVERSIDADE ESTADUAL DE FEIRA DE SANTANA

Abstract

Several species of medicinal plants generally contain in their composition volatile compounds. In general, these organic molecules act as repellents or attractive of pollinating insects. The aim of this work was to prospect new attractive compounds for Aedes aegypti L. through the Odorant Binding Protein (AeagOBP1). The AeagOBP1 structure was downloaded from PDB Database, with access code 3K1E considering the organism, resolution (1.85) and R-value (0.212). Structures of isolated compounds from semi arid plants were drawn using Marvin Sketch (Chemaxon). After, we verified valences, structural erros, and then we saved all ligands in MOL2 format. For docking studies, all ligands were prepared using AutoDock tools and saved in PDBQT format. Furthermore, we defined the active stite region (gridbox) for AeagOBP1 and saved the coordinates. Molecular docking calculations were performed using AutoDock Vina. After evaluated each docking positions, and cosidering best affinity energy and ligand positing inside AeagOBP1 active pocket, we used PyMOL 1.7 in order to save complexes in PDB format. 2D interaction maps for each complex were generated using Discovery Studio 4.0. In this work we tested 9 molecules deposited in the Semi Arid Molecules Database (SAM Database), hosted on the servers of the Bioinformatics and Computational Chemistry Lab (LBQC-UESB). SAM3814 ligand had best interaction with AeagOBP1 (-8.3 Kcal/mol). Standard ligands were tested for validation purposes: Carbon dioxide (-1.6 Kcal/mol), lactic acid (-3.2 Kcal/mol), octenol (-4.6 Kcal/mol) and 2-oxopentanoic acid (-4.2 Kcal/mol), however presented worst interaction energies when compared to SAM3814. This work demonstrated that natural volatile compounds isolated from Brazilian semi arid plants could act as new ligand prototypes in order to develop new attractive and/or repellent compounds for Aedes aegypti mosquito. On the other hand, this could be used as an new strategy for the controlling incidence of dengue, chikungunya and zika viruses.

Funding: Sem financiamento

Comparative analysis of the alternative splicing diversity in the human and mouse brain proteomes: preliminary results

Esdras Matheus da Silva[1], Thais Martins[1], Raphael Tavares da Silva[2], Fabio Passetti[3]

1 OSWALDO CRUZ INSTITUTE
2 UFMG
3 FIOCRUZ - IOC

Abstract

The alternative splicing of pre-mRNAs in eukaryotes can generate an extensive complexity of alternative protein variants from a given gene. Some mutations in the genome can cause malfunction of pre-mRNA splicing mechanism and, hence, generate protein variants with the potential to lead to neurodegenerative diseases. Model organisms, particularly, the mouse, are often used for the study of many pathologies because of ethical and methodological convenience. The mouse brain proteome has been investigated to improve the knowledge of the molecular aspects of neurodegenerative diseases. Currently, mass spectrometry (MS) is the most used technology for protein complex sample analysis. This methodology is based on the digestion of a given protein sample followed by their mass detection, as well as their identification by computational analysis with a traditional protein sequence repositories. However, some peptides are not identified by these traditional repositories because they comprise protein variants derived from alternative splicing events. As a solution, proteogenomic approaches have been used to identify these peptides that cannot be found in convencional repositories. In short, this strategy consists in using genome or transcriptome data to build customized protein repositories. Here, we investigated the diversity of protein variants formed by alternative splicing in the human and mouse brain, by using MS public data and two customized protein sequence repositories, one for each species, created by our group. First, we detected alternative splicing variants in complete reference mRNA (RefSeq) data along with ESTs through a methodology developed by our group called ternary matrices. Second, after an in silico translation, the predicted proteins were in silico digested and the resulting peptides were selected if they were not comprise in any sequence from Uniprot/SwissProt database. Third, customized repositories were created based on the union of selected peptides and Uniprot/SwissProt protein sequences. Forth, we made proteomic analysis using these repositories and comparative analysis of results from both species in order to identify the expression of orthologous genes at the protein level. The customized repository for humans had 20,150 canonical sequences and 204,294 nonredundant peptides from protein variants formed by alternative splicing. The mouse repository had 16,888 canonical sequences and 156,889 nonredundant peptides from protein variants formed by alternative splicing. In the MS experiments of the brain's corpus callosum for both species we identified 1,040 orthologous genes expressing canonical proteins and 5 orthologous genes expressing protein variants formed by alternative splicing. We believe that this study can contribute to the better understanding of the alternative splicing profile that can be found in both human and mouse's brains. Financial support: CAPES, FAPERJ and Fiocruz.

Analysis of splice variants in the proteome of Alzheimer's disease: preliminary results

Thais Martins[1], Esdras Matheus da Silva[1], Raphael Tavares da Silva[2], Fabio Passetti[3]

1 OSWALDO CRUZ INSTITUTE
2 UFMG
3 FIOCRUZ - IOC

Abstract

Alzheimer's disease, Parkinson's disease and prion disease are the most common neurodegenerative diseases, affecting millions of people worldwide. Currently there is no cure or preventive therapy or quick diagnosis for any of these pathological conditions, but in all of them, abnormal accumulations of protein aggregates occur in the brain. These pathologies have been correlated to altered proteins which can be derived from alternative splicing in the pre-mRNA. Therefore, the discovery of novel protein isoforms is an important strategy to identify new biomarkers for diagnosis, potential therapeutic targets or monitoring the development of each illness. Mass spectrometry data of cerebrospinal fluid (CSF) and brain tissue from patients with Alzheimer's disease were obtained from public databases to identify the protein profile and expression of protein variants generated by alternative splicing. In the analysis of CSF data, 9 common splice variants were identified between data from patients with Alzheimer's and control patients. In addition, 4 splice variants were unique to patients with Alzheimer's disease and the canonical proteins of these genes were directly correlated with the disease as described in the literature. In the analysis of brain tissue data, we identified 16 splice variants unique to patients with Alzheimer's, which 9 canonical proteins of these genes were directly correlated this disease according to the literature. From this analysis, most alternative splicing isoforms have been identified based on proteotypic peptides which were located at junctions from non consecutive exons. The study of exclusively expressed isoforms is an important strategy for the identification of new biomarkers for the monitoring of neurodegenerative diseases.

Funding: FAPERJ, CAPES and Fiocruz

Evaluation of differentially expressed proteins during Leishmania major infection in murine macrophages lacking nitric oxide synthase

Victor Hugo Toledo[1], Djalma de Souza Lima Junior[1], Lívia Rosa Fernandes[2], Giuseppe Palmisano[2], Luiza A. Castro-jorge[1], Dario Simões Zamboni[1]

1 FACULDADE DE MEDICINA DE RIBEIRÃO PRETO - USP
2 INSTITUTO DE CIÊNCIAS BIOMÉDICAS - USP

Abstract

Leishmaniasis is a neglected tropical disease that can have 3 different presentations, cutaneous, mucocutaneous, or visceral leishmaniasis. It is caused by a diverse group of protozoan parasites, Leishmania, and is transmitted by certain types of sandflies. It is estimated that 1.5 million people are infected each year in more than 98 countries where the disease is endemic. Until now, vaccination and drug therapy have failed to control the disease. The main mechanisms responsible for controlling Leishmania replication involves the production of nitric oxide (NO), generated by inducible NO synthase (iNOS) following activation by IFN? and also reactive oxygen species (ROS), generated by the respiratory burst. Hence, studies evaluating changes in protein expression after Leishmania major infection in wild type macrophages, and macrophages lacking or superexpressing iNOS could help in the discovery of novel targets for the control of L. major. In order to identify proteins differentially expressed (DEPs) related to these functions, we analyzed protein extracts from C57BL/6J bone marrow derived macrophages (BMDMs) infected or not with Leishmania major and iNOS-/- BMDMs, using mass spectometry. Differential regulated proteins were selected based on several statistical analyses (t-test, LIMMA, ROTS and SAM), performed using RStudio and relevant packages, as to combine results from several sources and to choose the most suitable method to improve the confidence. DEPs were then submitted to biological network analyses using Enrichment Map and Gene Ontology related tools (such as g:Profiler and DAVID) to define enriched functionally related genes. In addition, we evaluated protein-protein physical and functional interactions with STRING database and also pathway abundances through Ingenuity Pathway Knowledge Base to improve these results. Thereby, we identified proteins differentially modulated during L. major infection course, which allowed us to define important altered biological processes, such as early endosome to late endosome transport. We expect our results to widen the understanding of the infection control and to unravel new information for further studies.

Funding: Nenhum

In silico study of a new Brazilian semi arid compound with possible IKK-β inhibitory action

Wagner Rodrigues de Assis Soares[1], Thaís Almeida de Menezes[2], Bruno Silva Andrade[1]

1 UNIVERSIDADE ESTADUAL DO SUDOESTE DA BAHIA
2 UNIVERSIDADE ESTADUAL DE FEIRA DE SANTANA

Abstract

The family of Nuclear Transcription Factors (NF-kB) is among one of inflammatory central regulators, in innate and adaptive immunity. The enzyme IKK-β is responsible for the phosphorylation of NF-?B by modulating the transcription response of genes encoding proteins that participate in the immune and inflammatory response, cell adhesion, growth control and protection against apoptosis. In this study we tested several ligands isolated from Brazilian semi arid plants, in order to evaluated which best complexes with IKK-β structure. The enzyme structure was downloaded from PDB database, 4KIK, considering best resolution (2.83) and R-value (0.236). Ligand structures of were drawn in Marvin Sketch (Chemaxon), and deposited in the semi arid Molecules Database (SAM Database), hosted in the Bioinformatics and Computational Chemistry Lab (LBQC-UESB). After, we verified valences, structural errors, and saved all structures in MOL2 format. Additionally, all ligands were prepared for docking calculations, using AutoDock Tools and saved in PDBQT format. Furthermore, the active site of IKK-β was defined (gridbox) and all coordinates were recorded. Molecular docking calculations were performed by AutoDock Vina, searching nine different docking positions and following the manual recomendations. In order to select the best IKK-β-ligand complex, we consider the best energy value, as well as the ligand position inside the active site. PyMOL 1.7 was used to evaluate complexes and save them in PDB format. 2D interaction maps for each best complex were generated using Discovery Studio 4.0. The structures of the two compounds were designed in OSIRIS Property explorer software to calculate theoretical values of solubility (cLogP), hydrogen bonding donors (HBD), hydrogen bond acceptors (HBA), molecular weight (PM), Polar Surface Area (PSA), Drug-likeness and Drug-Score properties. These compounds were evaluated according to Lipinski Rule 5 for oral bioavailability. From the molecules deposited in the SAM database, the ligand SAM0530 showed a similar interaction with IKK-β, when compared to the classical inhibitor Staurosporine, both with an affinity energy of -11.0 Kcal/mol. The 2D interaction map shows the most molecular interactions as Van der Walls forces. In addition, the compound SAM0530 presented physical-chemical parameters that indicate a good oral bioavailability, not demonstrating toxicity in silico prediction. This work demonstrates that the Brazilian semi arid region can provide new chemical structures with potential for inhibition of IKK-β as an important molecular target of medical interest, since the dysregulation of NFk-β contributes to numerous inflammatory pathologies, as asthma, arthritis , cancer, diabetes, AIDS and inflammatory bowel disease.

Funding: Universidade Estadual do Sudoeste da Bahia

6 | RNA and Transcriptomics

IN SILICO IDENTIFICATION, CHARACTERIZATION AND PHYLOGENETIC ANALYSIS OF miRNAs IN WILD PEPPER

Ailton Pereira da Costa Filho[1], Monize Angela de Andrade[1], Laurence Rodrigues do Amaral[1], Matheus de Souza Gomes[1]

1 UFU

Abstract

Capsicum annuum var. Glabriusculum is a species of wild pepper with perennial and woody growth that due to its organoleptic characteristics is used in food as a flavoring. It is an important source of germplasm for the Capsicum genus, especially when used as a source of genes for resistance to disease. Due to human invasion, inadequate harvests and environmental degradation, their survival is threatened. Recently, the plant transcriptome has received attention from the scientific community to identify which miRNAs are regulating gene expression. The miRNAs are a class of small non-coding RNAs which length ranges from 20 to 24 nucleotides, and perform regulatory function in the organism. This class of small RNAs is involved in several biological functions, such as cell proliferation, apoptosis, and stress response. The objective of this work was to identify, characterize and analyze phylogenetically, putative miRNAs of Capsicum annuum var. glabriusculum and their orthologs. We searched for the probable mature miRNAs and precursors using miRBase. Pre-miRNAs were characterized as to their structural and thermodynamic characteristics. The conservation and alignment analyzes were performed using ClustalX 2.1 and RNAalifold. The secondary structures of the pre-miRNAs were obtained by RNAfold. Phylogenetic analysis of C. annuum var. glabriusculum pre-miRNAs and their orthologs using MEGA v5.2 was also performed. From the analysis, 101 putative miRNA families were obtained, and the families MIR-160, MIR-162, MIR-164, MIR-390, MIR-393 and MIR-828 showed high conservation in Solanaceae. It is worth mentioning that the targets of these families were described. The phylogenetic analysis of these miRNA families showed high conservation within their families and the phylogenetic distribution corroborated with the plant life tree. When comparing the secondary structures of the orthologs with the precursors it was evident that the pre-miRNAs are also conserved, mainly within the family Solanaceae. Thus, the obtained results amplify the understanding of the miRNA pathway in wild Capsicum annuum var. glabriusculum opening space for new inquiries regarding the regulation of gene expression in this species.

Funding: FAPEMIG, CNPq, UFU and CAPES

Association of Hfq/LSm protein with insertion sequence-derived RNAs is a prevalent phenomenon in prokaryotes

Alan Péricles Rodrigues Lorenzetti[1], Lívia S. Zaramela[2], Joao Paulo Pereira de Almeida[1], José Vicente Gomes-filho[1], Ricardo Zorzetto Nicoliello Vêncio[1], Tie Koide[1]

1 USP
2 UNIVERSITY OF CALIFORNIA SAN DIEGO

Abstract

Insertion sequences (IS) are mobile genetic elements present in most of prokaryotes. Their mobilization is known to contribute to the genetic variability of host genomes, usually promoting structural variation, disruption of genes and alteration of the transcription profile. These modifications can increase an organism's fitness in some circumstances, but neutral and deleterious effects are still more frequent. To avoid damaging events, transposition rates are kept low by different mechanisms, including the translational repression of transposase mRNA by its association with RNA binding proteins (RBPs) and/or antisense RNAs (asRNAs). In Bacteria (Salmonella enterica), the asRNA art200 binds to the transposase mRNA of an element from the IS200/IS605 family and can form a ternary complex with Hfq protein to prevent translation. In Archaea (Haloferax volcanii), LSm protein, the Hfq homologue, also have been found attached to sRNAs complementarily to a transposase mRNA in a RIP-Chip experiment, pointing out for a translational repression-based transposition regulation mechanism similar to the aforementioned. We analyzed RIP-Seq data for Halobacterium salinarum NRC-1 and found out that several IS-derived RNAs also bind to LSm in this organism. In addition, we point out that this protein is mainly associated with AU-rich RNAs in vivo, in accordance with results previously reported for H. volcanii in vitro assays. Furthermore, we have analyzed public data for several bacteria to find out that the association of LSm/Hfq with IS-derived RNAs is a prevalent phenomenon in prokaryotes. Now we are investigating whether the absence of LSm protein impact the rate of transposition in our model organism, and the results should guide our next steps to elucidate this mechanism in Archaea.

Funding: FAPESP and CAPES

Unraveling the lincRNA transcriptome of the mice olfactory system

Antônio Pedro de Castello Branco da Rocha Camargo[1], Marcelo Falsarella Carazzolle[2], Fabio Papes[1]

1 UNICAMP
2 BIOLOGY INSTITUTE - UNICAMP, NATIONAL CENTER FOR HIGH PERFORMANCE COMPUTING

Abstract

The olfactory system is a sensory system capable of detecting environmental chemical cues, leading to the sensation of an odor and/or behavioral and endocrine changes. In order to perform these functions, this system comprises two organs, the main olfactory epithelium (MOE) and the vomeronasal organ (VNO), found in the nasal cavity of mammals. The MOE detects odors and initiates their corresponding neural pathways, whereas the VNO detects intra and inter-species stimuli and starts inate behaviour, such as sexual, agressive and social.

Recently, a huge variety of long non-coding RNA (lncRNAs) has been discovered in several tissues, regulating gene expression and development. Given the unique properties of the process by which genes coding for MOE and VNO receptors are regulated, we hypothesize that lncRNAs might be involved in such regulation. In order to unveil intergenic lncRNAs (lincRNAs) that could be participating in the differentiation of the olfactory neurons, we've developed a pipeline to identify and functionally annotate lincRNAs preferentially expressed in these organs.

In order to identify new non-coding transcripts, a new lncRNA predictor was developed using cutting edge machine learning algorithms and techniques. This predictor classifies transcripts into coding or non-coding using several numerical descriptors described in the literature and others developed in this work that achieves a better classification quality than any published tool in the literature.

Using public RNA-Seq libraries from eight tissues, including the VNO and MOE, we've constructed a new mice transcriptome, which was used by the new lncRNA predictor to detect non-coding RNAs. The expression of the transcripts was quantified in all the libraries and the lincRNAs with olfactory-specific expression patterns were selected using a method based on cosine similarity.

In order to find olfactory-specific lincRNAs with interesting expression patterns, that could indicate some kind of biological function, statistical tests were performed to find transcripts that are differentially expressed between different conditions in the olfactory organs. We've found 173 lincRNAs that are differentially expressed between adult and newborn mice in the olfactory organs. Moreover, a total of 93 public MOE single-cell RNA-Seq libraries were quantified and ordered in a developmental trajectory so that we could find 131 lincRNAs whose expression changes as a function of cell differentiation.

Finally, in order to infer possible biological functions of the lincRNAs, a weighted correlation network analysis was done using the MOE single-cell expression data and the clusters containing lincRNAs were submitted for GO enrichment analysis.

Funding: FAPESP

In silico molecular subtype-specific drug targets prospection by integrating colorectal cancer and tumor-derived cell lines data

Cristóvão Antunes de Lanna[1], Nicole Scherer[2], Luís Felipe Ribeiro Pinto[3], Mariana Boroni[2]

1 LABORATÓRIO DE BIOINFORMÁTICA E BIOLOGIA COMPUTACIONAL [LBBC/INCA]
2 LABORATÓRIO DE BIOINFORMÁTICA E BIOLOGIA COMPUTACIONAL
3 PROGRAMA DE CARCINOGÊNESE MOLECULAR

Abstract

Colorectal cancer (CRC) is the third most prevalent carcinoma in the world. The molecular basis of CRC progression is well known and the vast amount of available data has allowed the classification of CRC tumors into molecular subtypes. However, many classification systems have been developed independently and they are generally inconsistent. Recently, the Colorectal Cancer Subtyping Consortium (CRCSC), an initiative involving several independent groups, has identified four consensus molecular subtypes (CMS) based on gene expression data from more than 4,000 primary tumor samples. Since cell lines are frequently used as in vitro tumor models, our study aims to classify CRC cell lines into their respective CMS and use them as platform to discovery/validate potentially new drug targets. In the present study we analysed a total of 155 CRC-derived cell lines using both Loess-normalized microarray data obtained from the GEO database and RNA-seq data from the SRA database. For RNA-seq data, after quality control and low-quality bases removal using FastQC and Trimmomatic, respectively, reads were aligned to the human genome and sorted by coordinate using STAR. After that, mapped reads for each gene were quantified and gene expression levels were normalized using RSEM. Additionally, raw RNA-seq aligned read count data for The Cancer Genome Atlas (TCGA) CRC primary tumor samples was downloaded using the TCGABiolinks tool and normalized using the DESeq2 tool, both written in R. Normalized counts from all sample sets were used to classify each sample using the CMSClassifier tool, written in R and developed by CRCSC. This tool allows us to make a molecular classification of CRC samples using high-throughput expression data (RNA-seq or microarray) based on the random forest method. Mutation Allele Frequency (MAF) files were also obtained from TCGA and separated by the samples' CMS classification. A list containing overlapping frequently mutated and differentially expressed genes will be used for metabolic pathway enrichment of each CMS using tools such as MetaCore, David and Reactome. Pathway components will be used for building interaction networks using String. Features containing the most connections will be used as candidates for novel drug target discovery using the Integrity database. Identified targets will then be compared with CMS representative cell lines' targets identified in the CancerRXGene database in order to compare cell lines' and primary tumors' potential treatment responses. First results include sample classification, which for the cell lines is as follows: CMS1, 42 cell lines (27%); CMS2, 62 (40%); CMS3, 10 (6%); CMS4, 2 (1%); undetermined, 39 (25%). For TCGA samples, classification distribution is: CMS1, 83 samples (13%); CMS2, 349 (54%); CMS3, 51 (8%); CMS4, 60 (9%); undetermined, 104 (16%).

PRELIMINARY ANALYSIS OF miRNAs IN THE GENOME OF Citrus sinensis

Douglas Santana[1]

1 UFU

Abstract

The production of citrus is a highlight in the Brazilian agroindustry, since Brazil is responsible for 60% of the world production of orange juice and its main exporter world-wide. Among the main citrus fruits stands out the orange, from Citrus sinensis species. Citrus fruits generally have both nutritional and medicinal properties. Several sorts of diseases affect the citrus crop causing serious damage to the economy. Nowadays, these issues are being addressed with highly polluting pesticides, and the damage generated due to its usage for both the consumer and the environment is severely alarming. Although there is plenty information in scientific literature, such as information about the biology of the Citrus spp. species, little has been done about the molecules involved in the regulation of gene expression in these organisms. A better understanding of these silencing pathways and their effector-molecules could help elucidate mechanisms that are not harmful to health and the environment for the control of pathogenic microorganisms and diseases caused by them. MicroRNAs (miRNAs) have been prominent among small non-coding RNAs because of their role in the regulation of gene expression. Thus, the objective of this work was to identify and characterize using in silico approaches miRNAs and their precursors in the genome of C. sinensis. We used an optimized algorithm with several filters based on structural and thermodynamic characteristics of conserved miRNAs. The RNAfold program was used to predict the secondary structure of the miRNA precursors and the RNALalifold and ClustalX 2.1 programs were used to generate multiple alignments to verify similarity with their ortholog precursors from other plant species. We identified 126 precursors of miRNAs, 177 mature miRNAs and 42 families in C. sinensis. Among the families, we emphasized the following miRNAs: miR156 / 157, miR390, miR2111a, miR3951 and miR3954. The miRNAs showed conservation at both the primary and secondary levels of structure. The phylogenetic distribution of C. sinensis miRNAs corroborated with the Tree of Life due to their results in phylogenetic analysis through the software Mega version 5.2. The results obtained increased the knowledge of regulation of gene expression in C. sinensis providing new challenges for the search of technologies for the control of pathogenic insects and microorganisms and diseases they can cause in Citrus spp. species.

Funding: CNPq, FAPEMIG, INCTV

The assessment of the impact of small deletions within human protein domains using transcriptome data: a pilot study in lung cancer

Fernanda Cristina Medeiros de Oliveira[1], Gabriel Wajnberg[2], Fabio Passetti[3]

1 FIOCRUZ-IOC
2 FIOCRUZ - IOC

Abstract

Deletions are examples of polymorphisms that can alter the protein sequence encoded by genes. These changes within the amino acid sequence can be connected with numerous human pathologies, such as cancer. Lung cancer has the highest worldwide incidence of all tumors, with an increase of 2% per year. For this reason, there is an interest to find new methods for diagnosis and treatments. High-throughput sequencing generates a large amount of genome or transcriptome data in a short time when compared to other sequencing methodologies. High-throughput sequencing can be used to identify new polymorphisms, such as deletions in the human genome. In this study, we used RNA-Seq data from six lung adenocarcinoma patients available at the Sequence Read Archive database (study ID SRP012656). We identified 2,388 protein domains affected: 1,137 in control tissue adjacent to the tumor, 729 in tumor samples and 522 in both control and tumor samples. We identified deletions in protein domains with high probability to be associated with cancer biology, such as deletions found in the genes SASH1, GRINA, TP53BP2, RAVER1 and NCOR2, which share the same protein domain termed large tegument protein UL36. Changes in this domain may be decisive to the development of lung cancer, because SASH1 plays an important role as a tumor suppressor in lung cancer. We also identified different altered domains in the TP53BP2 gene, such as ankyrin repeats Ank and Ank 2. The p53BP2 protein binds to the p53 tumor suppressor by these ankyrin repeats to increase its DNA binding activity. We identified, through this work, changes in amino acid sequences caused by small genomic deletions up to 100 nucleotides in length that affect protein domains of proteins previously associated with lung cancer. These new findings may be useful for new studies related to the identification of new biomarkers for diagnosis and new therapeutic targets.

Funding: CAPES, FIOCRUZ, FAPERJ, PIBITI/CNPq

CHARACTERIZATION AND INDENTIFICATION OF MATURE miRNAs AND THEIR PRECURSORS IN THE GENOME OF CULTIVATED PEPPER

Fernando Augusto Corrêa Queiroz Cançado[1], Monize Angela de Andrade[1], Laurence Rodrigues do Amaral[1], Matheus de Souza Gomes[1]

1 UFU

Abstract

The cultivated pepper Capsicum annuum L. (Zunla-1) is a plant of Solanaceae's family, one of the crops with the largest cultivated area of Brazil according to CEAGESP, and one of the most consumed by humans. Its high nutritional, sensorial and aesthetic value for different foods worldwide are directly proportional to its economic and cultural importance. Despite the significance of this species, the knowledge about the gene regulation is still very scarce. A class of microRNAs (miRNAs) is considered the main class of small non-coding RNAs with approximately 19 to 25 nucleotides. They regulate the expression of messenger RNA (mRNAs) into cells, inhibiting their translation process. In cells, the miRNAs play several roles, including development regulation, defense, response to stress and control of cell proliferation. Therefore, this work objective was identifying and characterizing mature microRNAs and their precursors in the genome of C. annuum L.(Zunla-1). The precursors and mature miRNAs were identified using an optimized algorithm based on the conserved characteristics of miRNAs. The ClustalX 2.0 and RNAalifold programs were used to generate alignment while the RNAfold program was used to predict the secondary structure of the precursor. The Phylogenetic analysis was performed in the Mega5.2 software by the Kimura 2 parameters. About 91 families of miRNAs scattered in the genome, among them miR160, miR162, miR164, miR393 and miR828 were identified and characterized. Of these 5 families investigated we've found 12 real precursors with conservation at primary and secondary level. It was observed in C. annuum L.(Zunla-1) that families such as miR160 and miR828, showed miRNAs that were evolutionarily distant from other organisms in the Solanaceae family, leading to speculation that there is in fact evolutionary distant or missing information in the database. While in the other families it can be observed that there is evolutionary conservation among them. The miR160 family has already been described as regulating the response factors to auxin, a hormone involved in the regulation of plant cell growth, demonstrating the importance of these small RNAs in the organism. This study will open new challenges and new perspectives to understand better the biology and the genome of pepper.

Funding: FAPEMIG, UFU, CNPq and CAPES

HIGH-THROUGHPUT SEQUENCING AND DE NOVO ASSEMBLY OF TRANSCRIPTOME OF Vigna unguiculata UPON VIRAL INFECTION

Flavia Figueira Aburjaile[1], João Pacifico Bezerra Neto[1], Bruna Piereck Moura[1], José Ribamar Costa Ferreira-neto[1], Ana Maria Benko-iseppon[1]

1 UFPE, CENTER OF BIOLOGICAL SCIENCES, GENETICS DEPT

Abstract

Plants are often submitted to adverse environmental conditions, such abiotic and biotic stresses. According to the Food and Agriculture Organization of the United Nations, biotic stresses are responsible for diseases leading to 32 - 42% of productivity decrease. To neutralize infections, plants first recognize the invading pathogens by quickly and efficiently activating molecular mechanisms. During infectious process, the plant response involves changes at physiological, biochemical and molecular level, through activation of specific gene expression programs. In this context, the study of the transcriptome is an excellent alternative for identification of genes involved in plant-pathogen interaction, mainly for species considered as non-models like cowpea bean (V. unguiculata). Our transcriptome assembly was generated from, 12 libraries obtained for two different cowpea cultivars (IT85F-2687 and BR14-Mulato) submited to Cowpea Severe Mosaic Virus (CpSMV) and Cowpea Aphid-born Mosaic Virus (CABMV), respectively. The libraries were generated with three biological replicates for each treatment and control group, allowing us to decode its molecular behavior towards the pathogen in question. The total RNA from all samples described above was sequenced on Illumina platform. The data were processed according to the following steps: (1) analysis of data quality, (2) assembly "de novo", (3) annotation and (4) statistical analysis of differentially expressed genes, according to the MUGQIC Pipeline protocol. These sequences were divided into structural or functional categories, according to the gene family of interest. After quality analysis, more than 350 milions were assembled, and returned 149.288 transcripts from 72.140 genes, presenting a mean of 1.289 bp of length and a N50 of 1.981. Our annotation aligned with Uniprot and Pfam sequences return 54.8% and 40.9% of results, respectively representing 81.822 transcripts. Additionally, we performed gene ontology enrichment returning annotation for 76.089 transcripts (50.97%). This assembly will be an important resource to improve our understanding of main mechanisms that help cowpea cultivars to tolerate virus infection.

Funding: CAPES, CNPQ, FACEPE.

EVALUATING THE COWPEA DEHYDRATION STRESS TOLERANCE BASED ON INOSITOL AND RAPHINOSIS PATHWAYS

João Pacifico Bezerra Neto[1], Flavia Figueira Aburjaile[1], José Ribamar Costa Ferreira-neto[1], Ana Maria Benko-iseppon[1], Mg Santos[2]

1 UFPE, CENTER OF BIOLOGICAL SCIENCES, GENETICS DEPT
2 UFPE, BOTANY DEPT, PLANT PHYSIOLOGY LABORATORY

Abstract

Plants evolved to survive in environments that often impose adverse conditions, such as abiotic and biotic stresses. They developed several survival mechanisms that enable the detection of environmental changes, as well as induction of specific responses to imposed stress conditions. Cowpea is one of the most important food and forage legumes in north- and northeastern Brazilian regions and its ability to survive under environmental pressure make it an ideal crop model to study the molecular mechanisms of drought tolerance. In this context, the identification and characterization of inositol (Ins) and raphinosis (RFO) pathways genes was carried out for cowpea in their transcriptome by computational methods. The cowpea transcriptome was assembled from 453 milions of reads, resulting in more than 185,000 non-redundant transcripts, which include transcriptional variants, splicing products. Using seed sequences obtained from the Kyoto Encyclopedia of Genes and Genomes (KEGG) Pathway database 1.119 transcripts were obtained, 521 related to inositol pathway and 598 transcripts associated with raphinosis pathway. It was possible identify 31 KO numbers associated with the raffinose pathway, whereas 29 KO numbers were related to the Inositol pathway. Among all 1.119 transcripts, 468 gene ontology terms were obtained (238 for Ins and 230 for RFO), being reallocated with different enzymatic/metabolic activities that its members perform. For the RFO pathway, the most important biological processes comprise the metabolism of carbohydrates, galactose and raffinose, whereas for Ins we found phosphatidylinositol phosphorylation, lipid catabolism and inositol biosynthesis. Our data pointed out the importance of Ins and RFO availability for cowpea under dehydration, where many cellular processes require many members of both pathways, especially plants which use free Ins to synthesize essential compounds, including those involved in hormonal regulation and stress tolerance.

Funding: CAPES, CNPq, FACEPE.

Comparison of bioinformatics approaches to evaluate altered GO processes in in vivo and in vitro studies of antineoplastics of OPEN TG-GATEs online database

Giordano Bruno[1], André Luiz Molan[1], Jose Rybarczyk-filho[1]

1 UNESP

Abstract

Toxicogenomics is a promising field that has been continuously developed in the last decades. It's main goal is to study the toxic phenotypes of various chemicals in biologic systems with the aid of technologies used by omics sciences. The in sillico component of a toxicogenomics study, still presents many challenges in regards to what methods should be used, for example, to determine what biological functions are significantly alterated using gene expression data as a starting point. In this work we propose the comparison of two bioinformatics methodologies for the determination of biological processes (BP), molecular functions (MF) and celular components (CC). Gene expression data of three antineoplastics, Cyclophosphamide, Etoposide and Lomustine from the OPEN TG-GATEs online database will be used. The studies that will be used for comparison are the High dose and 24 hours cases of Homo sapiens in vitro, Rattus norvegicus in vitro and Rattus norvegicus in vivo. The first methodology will be functional enrichment of differentially expressed genes (DEGs) and the second method is an application of Shannon's normalized function of entropy to determine the relative ativity and diversity of groups of functionally associated genes (GFAGs), this method is aplied by the EntropyClusterGenes R package. The gene expression data used as input for both methodologies will be normalized using the same protocol, the RMA (Robust-Multiarray Average) which is a widely used normalization procedure for affymetryx Genechip microarrays data. In the first method genes that present p-value<0.05 and LogFC > 1 or LogFC < -1 will considered as DEGs, Bioconductor's R package 'topGO' will then use these DEGs as input to determine p-values for Gene Ontology (GO) processes (BP, MF and CC). The second method will group genes acording to their GO and then use the groups total expression levels along with a bootstraping statistic and FDR of 0,05 to determine adjusted p-values for said groups. In both methods, GO processes that presented p-values< 0,001 were considered as significantly altered. In general, the GFAG method detected more process than DEGs method, but there certain cases where the DEGs method detected more processes. However both methodologies showed similar behavior in regards to the number of detected processes across the drugs. The methodologies also were concordant in some of the results. In the etoposide case, for example, processes related to DNA replication and cell division were found to be significantly down regulated in both methodologies.

Funding: CNPq processo134467 / 2016-7

Computer-aided protocol to revisit the cDNA library from Lonomia obliqua caterpillar: Identification of structural motifs related to inflammatory processes

Jaqueline Mayara de Araujo[1], Milton Y. Nishiyama-jr[1], Flavio Lichtenstein[1], Kerly Fernanda Mesquita Pasqualoto[1], Ana Marisa Chudzinski-tavassi[1]

1 INSTITUTO BUTANTAN

Abstract

The skin contact with the bristles of the Lonomia obliqua caterpillar leads to poisoning, which is characterized by consumption coagulopathy and secondary fibrinolysis. These events may progress to hemorrhagic syndrome and, consequently, to death. It is well-known that the signaling pathways involved in inflammation and in the coagulation cascade are interrelated. Therefore, we can study such signaling mechanisms in diseases related to inflammatory processes. Osteoarthritis (OA), for instance, is one of these diseases, and is characterized by cartilage degeneration, accompanied by inflammation of the joints, pain and loss of physical functions. The current treatments for OA are limited and involve either pain relief, or total replacement of the joints, in the late stages of the disease. In this regard, the identification of important new molecular targets in the signaling pathways involved in inflammatory processes would be crucial for the development of new and more effective drug candidates to treat OA. The L. obliqua cDNA library analysis, using bioinformatics and computer-aided approaches, has provided the identification of structural motifs related to functions involved in inflammatory processes. The re-visitation protocol has allowed the functional annotation and reclassification of 1,503 transcripts from the cDNA library of L. obliqua. Twenty-nine predicted protein sequences related to inflammatory functions were identified. The findings allow us to propose five novel peptide constructions (new chemical entities, NCE) to be further synthesized and assayed in human chondrocyte models (experimental validation). Also, the novel peptide constructions will be used as molecular tools for the identification of new molecular targets related to inflammatory processes.

Funding: Fapesp

Genes and pathways modulated during Guillain-Barré Syndrome

Raulzito Fernandes Moreira[1], Paulo Ricardo Porfírio do Nascimento[2], Glória Regina de Góis Monteiro[3], Mario Emilio Teixeira Dourado Junior[3], Selma Maria Bezerra Jeronimo[3], João Paulo Matos Santos Lima[4]

1 PROGRAMA DE PÓS-GRADUAÇÃO EM BIOTECNOLOGIA DOS RECURSOS NATURAIS, UNIVERSIDADE FEDERAL DO CEARÁ
2 INSTITUTO DE MEDICINA TROPICAL DO RIO GRANDE DO NORTE, UFRN.
3 INSTITUTO DE MEDICINA TROPICAL DO RIO GRANDE DO NORTE, UFRN
4 PROGRAMA DE PÓS-GRADUAÇÃO EM BIOINFORMÁTICA, UFRN

Abstract

Guillain-Barré syndrome (GBS) is an acute polyradiculoneuropathy, monophasic, and since the eradication of poliomyelitis is the principal cause of paralysis in the world. This syndrome seems to have an autoimmune component characterized, in part, by molecular mimicry with production of antibodies that causes severe damage to peripheral nerves. Such damage results in symptoms, which include acute flaccid paralysis. About 30% of the cases need respiratory assistance. GBS cases is usually preceded by infection agents, such as Campylobacter jejuni and viral infections. It seems that the pathogenesis of C. jejuni in GBS is associated with anti-gangliosides antibodies, which cross react with gangliosides present in the nerve axolemma, mainly in the peripheral nerves. Hence, the present study aimed to analyze transcriptomic libraries from patients with GBS, diagnosed with the different subtypes (demyelinating, axonal and Miller-Fisher), in order to identify genes and key pathways related to GBS development and potential target for modulation. For this, 24 libraries were obtained from 12 Brazilian patients diagnosed with GBS (6 from demyelinating subtype, 4 from axonal form and 2 from Miller-Fisher form), in two distinct phases, symptomatic/acute phase and after complete recovery. The quality analysis, alignment, assembly and global gene expression were performed using FastQC, Bowtie, TopHat, Cufflinks (Cuffmerge) HTSeq and edger. Approximately 2000 genes were differentially expressed between symptomatic and the recovery phase ($p<0.01$ and log fold change 1.5). Transcript annotation based on GO and KEGG terms showed changes in expression of genes related to inflammation, as TNF signaling pathway, toll-like and NOD-like receptor signaling pathways. Also, pathways related to neurodegenerative, autoimmune and infectious diseases were enriched during symptomatic phase when compared to recovery phase. These results are in accordance to other previous studies and provide an overview of possible responses during the course of GBS.

Funding: CNPq, NIH, FUNCAP, CAPES.

Identification and characterization of miRNAs and their targets in cucumber genome

Júlia Silveira Queiroz[1], Núbia Carolina Pereira Silva[1], Laurence Rodrigues do Amaral[1], Matheus de Souza Gomes[1]

1 UFU

Abstract

Cucumber (Cucumis sativus L.) is one of the mainly vegetables of the world belonging to Curcubitaceae family, its production and quality are, normally, influenced by different biological and environmental factors. The culture shows problems related to biotic and abiotic stresses and a way to change the organism to present positive characteristics is through small RNAs, as microRNAs (miRNAs). miRNAs direct messenger RNAs (mRNAs) to transcriptional repression in different conditions and development stages. And despite the vast knowledge of cucumber biology, little is known about the regulation of miRNA expression. Thus, through optimized bioinformatics tools and robust algorithms, the identification and characterization of miRNA molecules, precursors and targets were performed. Precursors of miRNAs (pre-miRNAs) were predicted using an algorithm based on a series of structural and thermodynamic characteristic filters of conserved precursors. Through Phytozome database, sequences of the genome of C. sativus with hairpin structures or similarity with pre-miRNAs structures were accessed. After filtration of these sequences, additional analysis were performed to compare the putative miRNAs with their respective orthologs. The sequences were subjected to sequence alignments using the ClustalX 2.0 and RNAalifold and then undergo phylogenetic analysis conducted by Mega5.2. Finally, target prediction was performed through the psRNAtarget. We identified in our study, 130 pre-miRNAs and 197 mature sequences from 42 miRNAs families and 87 targets from 98 different sequences. All our findings showed great MFE, AMFE and MFEI values, indicating that the results found are possibly real. Furthermore, alignments and phylogenetic analysis showed miRNAs highly conserved, corroborating our results with literature. One of the most conserved families found was csa-miR160, which regulates the expression of Auxin Response Factor (ARF) genes, they may be responsible for regulating two other targets, such as "b3 DNA-binding domain" and "Ammonium Transporter Family". Thus, the results allow us to expand the study of miRNAs in cucumber, providing new challenges for understanding the biology of this organism.

Funding: FAPEMIG, CNPq, UFU and CAPES

High throughput sequencing of small RNAs in Biomphalaria glabrata

Laysa Gomes Portilho[1], Fábio Ribeiro Queiroz[1], Wander Jesus Jeremias[1], Elio Hideo Babá[1], Roberta Lima Caldeira[1], Laurence Rodrigues do Amaral[1], Matheus de Souza Gomes[1]

1 UFU

Abstract

Biomphalaria glabrata is a mollusc intermediate host of Schistosoma mansoni, one of the causative agents of schistosomiasis, which currently affects about 240 million people worldwide. The interaction between the parasite and the snail is controlled by some genes related to the susceptibility/resistance of the host and the infectivity of the worms. Small RNAs, <200 nt length, have been reported to perform fine and specific gene regulation in various organisms. MicroRNAs and piRNAs are two important classes, have the function of regulating the expression of messenger RNAs by complementarity of bases. These molecules can be identified by computational analysis and experimental approaches. A widely used experiment to identify miRNAs is the next-generation sequencing. Thus, the aim of this work was to identify and characterize mature miRNAs and piRNAs from smallRNA-seq (Adult snails) and identify respective miRNA precursors in the genome of B. glabrata using in silico analysis. Data from the Vectorbase database were used for mirDeep analysis and the algorithm developed by our research group. Using the RNAalifold and RNAfold programs, the structural and thermodynamic characteristics of the identified pre-miRNA sequences were analyzed. In addition, the multiple sequence alignment was performed using ClustalX2 and the phylogenetic analysis was accomplished using the MEGA5.2 software and neighbor-joining program. 94 conserved mature microRNAs were found in the smallRNA sequencing and 71 precursors related to these mature miRNAs were found in the snail genome. From these precursors, 22 were considered Protostome-specific and five were assessed as Mollusca-specific. Furthermore, the primary and secondary structures of bgl-miR-71, bgl-miR-137, bgl-miR-184, bgl-miR-281 and bgl-let-7 were characterized and showed high conservation when compared to their orthologs including the cluster miR-71/2. The study of B. glabrata miRNAs may help to clarify many of biological processes which are related to their respective gene target. In addition, these results will add new information to what is known about this class of small RNAs in animals. The identification and characterization of miRNAs and their precursors in the intermediate host of schistosomiasis will expand and facilitate searching for new information and strategies to enhance the approaches used currently in order to prevent this disease.

Funding: UFU, FAPEMIG, CNPq and CAPES

Comparative analysis of transcriptomes reveals the existence of genes with distinct profiles: overactive genes and gaussian genes

Lissur Azevedo Orsine[1], Glaura da Conceição Franco[2], José Miguel Ortega[3]

1 UFMG
2 DEPARTAMENTO DE ESTATÍSTICA, UFMG
3 UFMG. LABORATÓRIO DE BIODADOS

Abstract

Presently several experiments of characterization of transcription in different tissues have been conducted. Here we analyzed the expression profiles of five experiments: ENCODE, FANTOM5, GTEx, Illumina and Uhlen, comprising respectively, 13, 56, 53, 16 and 32 tissues. They express over 1e-6 TPM (transcripts per million), respectively, 42, 21, 57, 47 and 44 thousands of transcripts. We have noticed that some genes vary the expression level around a mean value. Applying a statistic test, we determined that 13%, 10%, 3%, 18% and 12% of genes pass a test for normal distribution, thus we refer to these as primarily Gaussian (pG) genes. We also observed that often a gene has a high expression in a couple of tissues, but the remaining ones show a normal distribution. Removing the outliers (obtained from the boxplot procedure), we depicted cases where in some tissues the gene is overactive (in relation to the term "overexpressed") and in the complementary tissues the gene behave as Gaussian. Therefore, in some tissues the gene was considered complementary Gaussian (cG) while in other tissues the gene was overactive. The percentage of total Gaussian genes, adding up pG and cG increased to 68%, 21%, 11%, 60% and 61%. Thus, a regulatory mechanism that produces an average expression is very common. Moreover, overactive genes may play an important role in the tissues where they are overactive. A third category of genes was depicted by analyzing those that did not surpass a threshold of expression in the third quartile, a metrics derived from the boxplot procedure. These genes were named Tissue Specific (TS), because the expression is low in at least 3/4 of the set and where they are expressed, they present sharp peaks. Respectively, the percentage was 35%, 22%, 48%, 27% and 36%. We noticed that after removing outliers, actually the complementary tissues in some cases may present a very low but Gaussian expression. The percentage of genes that are TS in some tissues and the complementary set of tissues show Gaussian expression, represent, respectively, 35%, 0%, 1%, 27% and 36% of the total genes. The broad variation of these proportions is probably due to the accuracy of determining expression near the background by the distinct methodologies. However, the expression around an average is a remarkably frequent feature rather than an exception. Further analysis of the composition of the reported categories will be presented.

Funding: CAPES, CNPq and FAPEMIG

Combining metagenomics and metatranscriptomics approaches for prospection of CAZymes of the lower termite Coptotermes gestroi

Luciana Souto Mofatto[1], João Paulo Lourenço Franco Cairo[2], Melline Fontes Noronha[3], Ana Maria Costa Leonardo[4], Fabio Marcio Squina[5], Gonçalo Amarante Guimarães Pereira[6], Marcelo Falsarella Carazzolle[7]

1 UNICAMP
2 LABORATÓRIO NACIONAL DE CIÊNCIA E TECNOLOGIA DO BIOETANOL, CENTRO NACIONAL DE PESQUISA EM ENERGIA E MATERIAIS, CAMPINAS - SP
3 CENTRO PLURIDISCIPLINAR DE PESQUISAS QUÍMICAS BIOLÓGICAS E AGRÍCOLAS, UNICAMP
4 DEPARTAMENTO DE CIÊNCIAS BIOLÓGICAS, INSTITUTO DE BIOCIÊNCIAS, UNESP
5 UNIVERSIDADE DE SOROCABA, PROGRAMA DE PROCESSOS TECNOLÓGICOS E AMBIENTAIS
6 BRAZILIAN BIOETHANOL SCIENCE AND TECHNOLOGY LABORATORY, BRAZILIAN CENTER FOR RESEARCH IN ENERGY AND MATERIALS, BIOLOGY INSTITUTE - UNICAMP
7 BIOLOGY INSTITUTE - UNICAMP, NATIONAL CENTER FOR HIGH PERFORMANCE COMPUTING/UNICAMP

Abstract

Termites are interesting insects to mining new efficient enzymes for biomass degradation (CAZymes - Carbohydrate activity enzymes). Because they live in symbiosis with bacteria, protozoa and fungus inside their guts, termites have the ability to degrade approximately 90% of plant-dry matter in tropical forest, converting lignocellulosic materials into fermentable sugars. In a previous study of our group, we performed bioinformatics analysis of Coptotermes gestroi genomic and transcriptomic data for prospecting CAZymes in this termite, including symbiont genes. The results identified few CAZymes from symbiont species, probably due to the low representation of these genomes in comparison with the genome of C. gestroi. In order to prospect more specific CAZymes from symbionts, a new approach was performed using the combination of metagenomic and metatranscriptomic analysis from C. gestroi gut. The main aim of this study is to compare metatranscriptomic data from insects submitted to different diets using metagenomic assembly as reference, mainly composed by symbiont sequences. For this purpose, we obtained RNA-seq data from five conditions: (1) sugarcane bagasse in natura; (2) sugarcane bagasse treated with phosphoric acid; (3) filter paper (cellulose); (4) filter paper (cellulose) and iron; (5) sugarcane bagasse treated with sodium chlorite and hydrochloric acid.

Transcriptional evaluation of induced pluripotent cells from patients with Cockayne syndrome after induction of DNA damage triggered by oxidative stress

Maira Rodrigues de Camargo Neves[1], Livia Luz Souza Nascimento[1], Alexandre Teixeira Vessoni[2], Carlos Frederico Martins Menck[1]

1 DEPARTMENT OF MICROBIOLOGY, INSTITUTE OF BIOMEDICAL SCIENCES, USP

2 DEPARTMENT OF MEDICINE, WASHINGTON UNIVERSITY IN SAINT LOUIS

Abstract

Cockayne Syndrome (CS) is characterized by symptoms related to premature ageing with severe involvement of the central nervous system. The molecular basis of the disease is related to deficiency in the transcription-coupled repair (TCR), mainly with mutations in the ERCC8 and ERCC6 genes (coding for CSA and CSB proteins, respectively). The phenotype of CS cells is presented as high sensitivity to ultraviolet (UV) light, causing DNA damage, which in turn prevents transcription recovery after irradiation. They are also more susceptible to DNA damage caused by oxidative stress, which maybe responsible for endogenous DNA lesions. Although it has been proposed that the CS transcriptional pattern following DNA lesions might be responsible for the cellular and clinical phenotype of patients, this pattern has not been investigated yet for stem cells. In the present work, we are investigating the transcription pattern through RNAseq in CS induced pluripotent stem cells (iPSCs) following DNA damage by oxidative stress. Preliminary tests for cell survival determination allowed the standardization of a Potassium bromate (KBrO3) concentration for DNA damage challenge experiments. Experiments were conducted on a wild type cell strain (F9048), and on a CSB mutant (GM10903, Coriell), both reprogrammed to iPS. Libraries were prepared for RNAseq with mRNA from both cell strains, extracted 24 h after KBrO3 and mock treatments. Sequencing was conducted on an Illumina NextSeq, with paired-end reads. The run yielded 637 million clusters, with an average of 52 million paired-end reads per sample. Data analysis was performed with the HISAT2-StringTie-Ballgown protocol (Tuxedo 2), against Ensembl GRCh38 genome. RSeQC was used for quality control and determination of median transcript integrity number (medTIN), and distribution of reads along each transcript to exclude library preparation bias. CS cells presented 109 differentially expressed genes, all observed exclusively on these cells following DNA damage challenge. Interestingly, only one gene (VLDLR-AS1) was identified as differentially expressed in wild type cells under the same treatment, suggesting CS cells are more sensitive to transcriptional variation after oxidative stress. An enrichment of GO terms for the regulation pathway of insulin growth factor (IGF) was found among the differentially expressed genes on CS cells, but not on wild type cells, corroborating previous findings. Furthermore, over half of the differentially expressed genes on CS are associated with the GO biological process "response to stimulus", mainly "response to stress". Five differentially expressed genes were classified as GO neuron projection regeneration (ULK1, SPP1, APOE, ADM, JUN), and possibly contribute to the severe nervous system involvement in the patients phenotype.

Finders keepers, nobody weepers! Unraveling novel genes in transcriptomes.

Marina Pupke Marone[1], Felipe Rodrigues da Silva[2]

1 INSTITUTE OF BIOLOGY, UNICAMP
2 EMBRAPA INFORMÁTICA AGROPECUÁRIA

Abstract

RNA-seq has become a standard procedure for measuring gene expression levels as it is a very sensitive and accurate tool. Studies involving RNA-seq generate a big amount of data and most of them are focused on finding the differentially expressed known genes, ignoring novel ones that might be present in the dataset. We are trying to devise an optimal pipeline to find novel genes in transcriptomes from organisms which have their genomes sequenced and several datasets available on the SRA database. In order to do this, we are going to use data from A. thaliana because of its importance as a model organism, making it easier to work with, added to the fact that there is a lot of data available. Among the 664 RNA-seq datasets found online for A. thaliana, we chose two very distinct to be analyzed. Both sets use the wild-type Col-0 ecotype, but the RNA of one of them was collected from the inflorescences, while the other was collected from the whole plant. Those sets were chosen to test whether RNA extracted from less complex tissues increase the chance of finding new genes. Transcriptomes sets were assembled using StringTie and Trinity in order to evaluate which one is the most appropriate for this task, considering the number of ESTs found and their performances. A transcript is deemed novel when it is described on our assembly but missing on the most recent genome annotation for that species. The presence of "old school" ESTs increases the confidence on the existence of the transcript.

Funding: CNPq

MiRNA, piRNA and snoRNA expression profile analysis in thyroid cancer subtypes

Mayla Abrahim Costa[1], Natasha Jorge[2], Fabio Passetti[2]

1 IOC/FIOCRUZ - RJ
2 FIOCRUZ - IOC

Abstract

Thyroid cancer is a public health problem and is considered the most common malignant tumor of the endocrine system. Different treatment strategies have been developed to improve patient's condition, including the identification of molecular markers. Over the years, non-coding RNAs (ncRNAs) have been identified as potential molecular markers capable of predicting therapeutic outcome. Studies show that small ncRNAs (sncRNAs), such as microRNAs (miRNAs), Piwi-interacting RNA (piRNA) and small nucleolar RNAs (snoRNAs) play important roles in cancer and response to treatment. In order to identify the miRNA, piRNA and snoRNAs constitutively and differentially expressed in samples of the different thyroid cancer subtypes, we analyzed small RNA high throughput sequencing data of thyroid carcinoma samples. We obtained normal and tumor samples of carcinoma papillary (PTC) (49 patients), carcinoma papillary, follicular variant, (PCF) (7 patients), and carcinoma papillary, columnar cell variant, (PCC) (3 patients), accounting for 118 paired samples available at the Genome Atlas Database. Forty six differentially expressed sncRNAs were obtained, of which 40 were miRNAs. Among them, 21 were detected as up regulated in tumor samples and 19 were down regulated. A total of 6 snoRNAs have been detected differentially expressed in all comparisons, 2 more expressed in tumor samples and 4 snoRNAs more expressed in control samples from the tissue adjacent to the tumor. We identified 34 constitutively expressed sncRNAs, including the piRNA hsa-piR-009294, in the 3 tumor subtypes. The integration of the differential expression and dispersion analysis revealed 3 miRNAs presenting similar expression pattern in tumor subtypes PCC and PTC when compared to the constitutive expression pattern in control and tumor samples of the PCF subtype. These results show that it was possible to detect sncRNAs differentially and constitutively expressed in samples of the different thyroid cancer subtypes. Our results corroborate those obtained by others and present novel findings, evidencing a viable alternative to search for novel potential molecular markers.

Funding: CAPES, FAPERJ and FIOCRUZ

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

Nayara Toledo[1], Raphael Tavares da Silva[1], Tiago Bruno Rezende de Castro[1], Glória Regina Franco[1], Andrea Mara Macedo[1], Carlos Renato[1], Égler Chiari[1], Neuza Antunes Rodrigues[1]

1 UFMG

Abstract

Chagas disease is a parasitic infection caused by the protozoan Trypanosoma cruzi. Even after 100 years of its description, the causes of the different clinical manifestations are not completely understood, although they certainly involve both parasite and host features. 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 contributes for changes in the differential tissue distribution of T. cruzi during infection were also found. RNA-Seq of mRNA extracted from BALB/c infected hearts (groups: JG, Col1.7G2 and an equivalent mixture of both strains) showed that Col1.7G2 was a strong activator of immune response genes, while JG effectively modulated the oxidative stress response and protein synthesis in the host. Curiously, the mixture-infected group showed both features simultaneously. 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 aim of the present study is to integrate mass spectrometry-derived proteomics from BALB/c infected hearts with the same T. cruzi strains and the above mentioned RNA-Seq data. We will investigate if the parasite can remodel the splicing pattern of the host and if this remodeling can influence on the disease development. For initial analyses, reads were mapped against mouse reference genome using the splice-aware aligner, STAR. Subsequently, full-length transcripts were reconstructed with Trinity and their quality was assessed by the Transrate software. Up to now, we performed only a de novo transcriptome assembly for the control group to adjust and to define the best parameters that provide an optimal assembly. Results reported by Transrate indicated that the k-mer length of 25 and a minimum k-mer coverage of six were the parameters which best performed the de novo transcriptome assembly. Our future steps include genome-guided transcriptome assembly, analysis of alternative splicing expression and correlation with proteins identified in mass spectrometry data.

Funding: CAPES e FAPEMIG

Unraveling the molecular profile of alternative transcripts through analysis of eCLIP and RNA-Seq data

Pedro Rodrigues Sousa da Cruz[1], Felipe Ciamponi[1], Katlin Massirer[1]

1 CBMEG - UNICAMP

Abstract

Genomic studies estimated that around 95% of human genes undergo alternative splicing and about 37% of them generate multiple protein isoforms, thus adding to the proteome complexity. Since pre-mRNA splicing is an essential process in mammalian cells, its failure can lead to overall or tissue-specific misregulation, possibly leading to diseases such as cancer. Although there has been a striking progress in uncovering the regulatory aspects of splicing networks, there is still relatively poor knowledge on the mechanisms that drive the occurrence of specific splicing events. RNA-biding proteins (RBPs) compose the main class of splicing regulators by binding to sets of mRNA-targets. We aim to assess the characteristics of the splicing site regions for those mRNA-targets for both control and RBPs individual knockdown cells in order to understand features that lead to specific events. To accomplish that, we built an analysis pipeline consisting of processing publically available enhanced CLIPseq (eCLIP) data from ENCODE; composing a reliable RBP-target list by filtering significant peaks; analyzing splicing patterns affected by these RBPs' knockdown from RNA-Seq data present in Genome Expression Omnibus (GEO; data also retrieved from ENCODE on the same cell lines) using FASTQC, trimmomatic, STAR, rMATS and R plotting functions; and finally building the profile of regions differentially used in splicing compared to the profile of general RBPs targets found in the first step. To perform the profiling we applied an algorithm developed by our group named BioFeatureFinder that gathers 5,498 features ranging from conservation data to physical characteristics as input and ranks them by significance. The eCLIP analysis showed the following splicing RBPs to be bound in splicing regions: TROVE2, PRPF8 (5' portion of the splicing region), SF3B4, SF3A3, U2AF1 and LARP7 (occupying the 3' portion). As for the profiling, BioFeatureFinder showed conservation to be a major aspect in RBPs target regions, additionally, these regions were found to be enriched for interactions with BUD13, SMNDC1 and EFTUD2, another splicing RBPs. Moreover, GC content and secondary structures were inferred to be important features shared by different targets under study.

Funding: CAPES, FAPESP

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

Priscila Grynberg[1], Larrisa A. Guimarães[1], Marcos Mota do Carmo Costa[1], Roberto Coiti Togawa[1], Ana Cristina M. Brasileiro[1], Patricia Messenberg Guimarães[1]

1 EMBRAPA RECURSOS GENÉTICOS E BIOTECNOLOGIA

Abstract

Plant-parasitic nematodes have a worldwide distribution. They are virtually able to infest any human-cultivated plant. Annual losses caused by nematodes on life-sustaining crops are estimated to exceed 14% of the production (approximately 65 billion €of loss worldwide). Previously studies were responsible for major advances in the identification of genes and mechanisms responsible for plants response to the Meloidogyne, the root-knot nematode. Meloidogyne spp. are obligate endoparasites that maintain a biotrophic relationship with their hosts. During the infection root cells are differentiated into specialized giant feeding cells through the releasing of effector proteins. However, despite the continuing efforts to identify new effectors and plant resistance mechanisms, studies have shown that the repertoire of both systems is limited. Recently, researchers published strong evidence that small RNAs from a phytopathogenic fungus act as effectors. These small RNAs hijack the host RNA interference (RNAi) machinery by binding to Arabidopsis Argonaute 1 (AGO1) and selectively silencing host immunity genes. These findings gave new insights on nematode-plant interaction as well as for the development of new control strategies through biotechnological methods. The goal of this work is to verify the possible role of Meloidogyne arenaria small RNAs (sRNA) as effectors by identifying, in Arachis stenosperma (peanut), downregulated target genes during the infection. A. stenosperma plants were infected with approximately 5,000 M. arenaria larvae in triplicate. Control and infected A. stenosperma roots were collected 3, 6 and 9 days post-infection. The infected samples were pooled. Six samples (3 controls, 3 infected) and two M. arenaria J2 small-RNA libraries were sequenced with technical replicates using Illumina HiSeq 2500 system. After adaptor and contaminats removal, reads were mapped against miRbase V21.0. One hundred and 151 different conserved miRNAs were counted for M. arenaria and A. stenosperma respectively. The unmapped reads were used as input for miRDeep-P, a plant microRNA prediction tool. The predicted miRNAs were confirmed by using miRDup softare . A total of 625 and 1271 new candidates were predicted for M. arenaria and A. stenosperma respectively. Next steps include the target prediction and validation by qRT-PCR.

Funding: FAP-DF, CNPq

Metalloproteinases diversity in the venom gland of Peruvian spider Loxosceles laeta revealed by transcriptome analysis

Raissa Medina Santos[1], Clara Guerra Duarte[2], Priscilla Alves de Aquino[1], Anderson Oliveira do Carmo[1], César Bonilla[3], Evanguedes Kalapothakis[1], Carlos Chavez-olortegui[1]

1 UFMG
2 FUNDAÇÃO EZEQUIEL DIAS
3 INSTITUTO NACIONAL DE SALUD

Abstract

Envenomation caused by spiders from Loxosceles genus (brown spiders) is a worldwide public health problem. Loxosceles their venom is composed of several toxins responsible for dermonecrotic, hemorrhagic and edema effects. In Peru, L. laeta is considered the most medical relevant species. A family of metalloproteases, also named astacin-like proteins, was described in Loxosceles venom with great importance for hemostatic disorders in natural or experimental envenomations. A new generation sequencing library of venom extracted from the Peruvian spider, L. laeta, was constructed for the first time using the TruSeq (TM) RNA Sample Prep Kit v3 Set A (Illumina) kit and the sequencing was performed on the MiSeq by the paired-end technique for identification of molecular diversity of metalloproteases toxins. In this work, we describe some of the identified metalloproteases enzymes with a high degree of identity (over 50%) with molecules from other Loxosceles spp spiders. Results obtained in this work represent the first landscape of components of a Peruvian spider venom gland, revealing the complexity of molecules expressed in this tissue, with great potential for future uses in medical and evolutionary studies.

Funding: FAPEMIG, CNPq, CAPES

TPP riboswitch analysis using molecular dynamic with different force fields

Rodrigo Bentes Kato[1], Jadson Claudio Belchior[1], Debora Antunes[2]

1 UFMG
2 INSTITUIÇÃO: INSTITUTO OSWALDO CRUZ - FIOCRUZ

Abstract

Riboswitch RNAs are important in bacterial metabolism and represent a promising class of antibiotic targets for treatment of infectious disease. Molecular dynamics simulations are used for interpreting experimental data and to predict new experiments. The present work is concerned to analyze the parametrizations most used in literature to describe force fields. A comparison is done between Charmm27 force field and others such as Amber99, AmberGS and Amber2014. Applications are carried out for tackling RNA molecules. The Gromacs (GROningen MAchine for Chemical Simulations) software are used to analyze RNAs structure and dynamic under a broad variety of patterns. In particular, Thiamin pyrophosphate (TPP) riboswitch was considered to evaluate strategies for studying parametrization of force fields. As it is well-known this RNA is important for regularization of gene expression through a variety of mechanisms in archaea, bacteria and eukaryotes. Main preliminary results are concerned with the preparation of RNA (2gdi.pdb) using a box of 7x4x3 angstrom, solvated with water TIP3 and neutralized with magnesium (Mg). The dynamics were carried out considering at this preliminary analysis up to 200 ns and the final analyzes were done using rmsd (root mean square deviation) for the trajectories. At this stage the outcome results demonstrated that AmberGS stabilized the RNA with the lowest rmsd, but closer to Amber99. In general, comparison using rmsd against others force fields showed: AmberGS 0.2182, Amber99 0.2312, Charmm27 0.5325 and Amber2014 0.6469. Therefore, AmberGS and Amber99 produced the best force fields for describing this particular RNA and it might be a good estimation to similar analysis for others systems and this is under investigation.

Funding: FAPEMIG, CAPES, CNPq

Annotation of transfer RNAs and microRNAs from Coffea canephora genome

Samara Mireza Correia de Lemos[1], Alexandre R. Paschoal[2], Douglas Silva Domingues[2]

1 UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
2 UTFPR - PPGBIOINFO

Abstract

Annotating plant genomes for non coding RNAs (ncRNAs) is helpful in the development of biotechnological products and plant breeding. Coffee is one of the most important commodities in agriculture and Brazil is the leading producer and second largest consumer market of coffee; however, research has mostly focused on identifying protein coding genes with few approaches addressing the non coding RNA component of coffee genome. We here used bioinformatic approaches to update the annotation of microRNAs and annotate transfer RNAs in the Robusta coffee (C. canephora) genome. Combining sequence similarity (BLASTN against ENSEMBL Plants database) and structural searches (Infernal/Rfam and tRNAscan - SE for transfer RNAs), we identified a set of 208 microRNA precursors and 663 transfer RNAs with their respective amino acids. A total of 144 microRNA precursors were identified for the first time in the present analysis: 122 using sequence similarity search and 22 using structural search. Sixty-four precursors were previously identified in a recent annotation; the majority, 63 were obtained in structural search. The proportion of transfer RNAs (tRNAs) was relatively similar to Populus trichocarpa and Vitis vinifera. The most common anti-codon is tRNA for methionine with 65 genes and the rarest is tRNA for tyrosine with only 14 copies. Our results represent an important improvement of the coffee ncRNA annotation, paving the way to further research on the contribution of post-transcriptional regulation to plant development and physiology. Future steps in this study include the annotation of other ncRNA classes and transcriptional support from small RNA sequencing data.

Funding: CAPES

A potential link between tuberculosis and lung cancer through non-coding RNAs

Sandeep Tiwari[1], Debmalya Barh[2], Ranjith N. Kumavath[3], Vasco A de C Azevedo[4]

1 1. INSTITUTE OF BIOLOGICAL SCIENCE, UFMG
2 LABORATÓRIO DE GENÉTICA CELULAR E MOLECULAR, DEPARTAMENTO DE BIOLOGIA GERAL, INSTITUTO DE CIÊNCIAS BIOLÓGICAS, UFMG, PAMPULHA
3 DEPARTMENT OF GENOMIC SCIENCES, SCHOOL OF BIOLOGICAL SCIENCES, CENTRAL UNIVERSITY OF KERALA, KASARAGOD, INDIA
4 UFMG

Abstract

Pulmonary tuberculosis caused by Mycobacterium and lung cancer are two major causes of deaths worldwide and the former increases the risk of developing lung cancer. However, the precise molecular mechanism of Mycobacterium associated increased risk of lung cancer is not entirely understood. Here, using in silico approaches, we show that hsa-mir-21 and M. tuberculosis sRNA_1096 and sRNA_1414 could play important roles in the pathogenesis of both these diseases. Further, we postulated a "Genetic remittance" hypothesis where these sRNAs may play important roles. The sRNA_1096 could be involved in tuberculosis through multiple infectious processes, and if transferred to the host, it may activate the TLR8 mediated pro-metastatic inflammatory pathway by acting as a ligand to TLR8 similar to the mir-21 leading to lung tumorigenesis and chemo-resistance. Analogous to SH3GL1, it may also regulate cell cycle. On the other hand, sRNA_1414 is probably involved in survivability and drug response of the pathogen. However, it may be a metastatic factor for lung cancer providing EPS8L1 and SORBS1 like functions upon remittance. Further, all these three non-coding RNAs are predicted to act in rifampicin resistance in Mycobacterium. Currently, we are applying robust bioinformatics strategies and conducting experimental validations to confirm our in-silico findings and hypothesis.

Funding: TWAS-CNPq , CAPES

Transcriptome profiles of Resistance Gene Analogs in Saccharum hybrid cultivar RB925345 in response to Sporisorium scitamineum infection

Sintia Almeida[1], Patricia Dayane Carvalho Schaker[1], Claudia Barros Monteiro-vitorello[1]

1 USP

Abstract

Sporisorium scitamineum is a biotrophic fungus responsible for the sugarcane smut, a worldwide spread disease. The disease is one of most harmful to the crop and occurs in all producing countries. The development of sugarcane smut symptoms depends on the interaction among environment, the sugarcane genotypes and the pathogen itself. RGAs (Resistance Gene Analogs) play a central role in recognising PAMPs, MAMPs, DAMPs or effectors from pathogens, triggering downstream signalling during plant disease resistance. RGAs comprise both cell surface pattern-recognition receptor (PRRs) and R-genes and can be grouped into several superfamilies based on the presence of a few structural motifs and conserved domains. For instance, both receptor like kinases (RLK) and membrane associated receptor-like proteins (RLP) are PRRs, while R-proteins are intracellular immune receptors mostly belonging to nucleotide-binding site-LRR (NBS-LRR). In general, the most prevalent R genes in plants are NBS-LRR, which are divided into two subclasses based on the presence of an N-terminal CC or TIR domain. To investigate expression profile and functional characterization of RGAs in sugarcane in response to smut fungus, a set of 16,219 transcripts from RB925345 susceptible variety was analyzed. RGAs were identified and categorized using PRGDB database. Differentially expressed (DE) RGAs were identified using a set of RNAseq data from smut-infected plants in two time points: 5 days after inoculation (DAI) and 200 DAI (after whip emission) using CLC Workbench V.8 (p-value < 0.05). To assess the functions of DEs, was performed a functional categorization was performed based on the KEGG database and Blast2go annotation. We identified 320 RGAs classified in four superfamilies: NBS (110) and TM-CC (29) containing proteins and membrane associated RLPs (39) and RLKs (142). Of them, 15 were differentially expressed (DE) at 5 DAI and 46 at 200 DAI. Lectin receptor-like kinases (LecRLKs) were abundant among the DEs, in which the extracellular lectin domain is known to bind to pathogen cell wall components. Genes related to effectors recognition belonging to the NBS family were also identified, such as the RPM1 disease resistance (R) protein homolog. These results show that comparative genomic analysis can help to identify host proteins related to pathogen recognition. Along with the analysis of differential expression we were able to determine those most responsive to the stimuli. Furthermore, the DE transcripts encoding RGAs may be used as molecular markers for resistance or susceptibility to smut disease.

Funding: CNPq and FAPESP

Transcriptome analysis of xylose and glucose co-fermentation by industrial engineered yeast for second generation bioethanol

Sheila Tiemi Nagamatsu[1], Luige Armando Llerena Calderon[2], Lucas Salera Parreiras[3], Bruna Tatsue Grichoswski Nakagawa[2], Angelica Martins Gomes[4], Gonçalo Amarante Guimarães Pereira[5], Marcelo Falsarella Carazzolle[6]

1 BRAZILIAN BIOETHANOL SCIENCE AND TECHNOLOGY LABORATORY), BRAZILIAN CENTER FOR RESEARCH IN ENERGY AND MATERIALS), BIOLOGY INSTITUTE - UNICAMP
2 BIOLOGY INSTITUTE - UNICAMP
3 BRAZILIAN BIOETHANOL SCIENCE AND TECHNOLOGY LABORATORY, BRAZILIAN CENTER FOR RESEARCH IN ENERGY AND MATERIALS, BIOLOGY INSTITUTE - UNICAMP
4 BRAZILIAN BIOETHANOL SCIENCE AND TECHNOLOGY LABORATORY, BRAZILIAN CENTER FOR RESEARCH IN ENERGY AND MATERIALS
5 BRAZILIAN BIOETHANOL SCIENCE AND TECHNOLOGY LABORATORY, BRAZILIAN CENTER FOR RESEARCH IN ENERGY AND MATERIALS, BIOLOGY INSTITUTE - UNICAMP
6 BIOLOGY INSTITUTE - UNICAMP, NATIONAL CENTER FOR HIGH PERFORMANCE COMPUTING/UNICAMP

Abstract

Second-generation (2G) ethanol is a promising technology which can increase production and reduce costs related to first-generation (1G) ethanol. Both process differ basically in the raw material for fermentative step, while 1G is based on fermentable sugars (glucose, fructose and sucrose) from sugarcane, 2G is based on deconstruction of biomass releasing fermentable sugars. This process generates non-fermentable sugars (mainly xylose) and inhibitors of yeast growth (acetic acid, furfural and HMF). To overcome these problems and increase yeast productivity in 2G ethanol production is essential select robustness microorganisms and perform genetic modifications to allow xylose consumption through insertion of endogenous xylose pathway genes, as xylose isomerase. Furthermore, approaches as evolutionary engineering can be used to improve some characteristics. Our previous work performed a comparative genomic analysis in genetically modified yeast followed by evolutionary adaptation for xylose consumption showing several point mutations and an increase of xylose isomerase genes during the evolution process. In this work we are showing a transcriptomic analysis from one parental (A) and two evolved strains studied before in xylose and glucose co-fermentation: one haploid strain (C) from intermediate round of evolution and the other, a diploid strain (E), from the final round of evolution. It was sequenced in biological duplicate and three different fermentation points for each strain, the first one with high glucose concentration and inhibition of xylose

Using CORAZON to investigate functionally and evolutionarily related coding and non-coding transcripts

Thaís de Almeida Ratis Ramos[1], Thaís Gaudencio[2], Vinicius Maracaja Coutinho[3], José Miguel Ortega[4]

1 UFRN
2 UFPB
3 UNIVERSIDAD MAYOR
4 UFMG. LABORATÓRIO DE BIODADOS.

Abstract

Machine learning is a subfield of computer science that developed from the study of pattern recognition and computational learning theories in artificial intelligence. These methods operate through the construction of a model based on the set of inputs, in order to make data predictions. Due to the large quantity of biological data generated in large-scale genomics and transcriptomics projects, an intense demand to use techniques provided by artificial intelligence, the usage of tools based on machine learning methods became widely used in bioinformatics. Unsupervised learning is the machine learning task of inferring a function to describe the hidden structure from unlabeled data. The inductor analyzes the examples provided and tries to determine if some of them can be grouped in any way, forming clusters. Here we developed an online tool calling CORAZON (Correlation Analyses Zipper Online) that include 3 unsupervised machine learning algorithms: Mean shift, K-means and Hierarchical with the bioinformatics purpose. Furthermore, we implemented 4 normalization methodologies: Transcripts Per Kilobase Million (TPM), base-2 log, instance normalization and normalization by the highest attribute value for each instance. Moreover, the user has a option to cluster the data removing each attribute to see the results, to observe the attributes influence. Here, we used our tool to study the coding and non-coding genes from Uhlen, Fantom and Encode databases. We found clusters well defined with genes of each of the two classes and analyzed biological processes determined by GO Enrichment Analysis and gene ages defined by Life Cycle Assessment (LCA). Normally, clusters with more codings are associated with cellular, metabolics, transports and systems development processes. Clusters with more non-codings are involved with detection of stimulus, sensory perception, immunological system, and digestion. We also observed that clusters with more than 80% of non-codings, more than 40% of their coding genes are recents appearing in mammalian class and the minority are from eukaryota class. Otherwise, clusters with more than 90% of coding genes, have more than 40% of them appeared in eukaryota and the minority from mammalian. Clusters without these criterias, have the majority of their coding genes arise from Eumetazoa. Therefore, the CORAZON tool can help in the large quantities analysis of genomic data, facilitating to comprehend the relations between these instances. In addition, as future work, it will be possible to understand the evolutionary history of these sequences and the associated transcription factors.

Funding: UFRN

Analysis of the role of an RNA binding protein in the control of gene expression in Trypanosoma cruzi epimastigotes

Wanessa Moreira Goes[1], Bruna Mattioly Valente[1], Edson Oliveira[1], Thaís Silva Tavares[1], Fabiano Sviatopolk Mirsky Pais[2], Caroline Leonel Vasconcelos de Campos[1], Santuza Maria Ribeiro Teixeira[3]

1 UFMG
2 CENTRO DE PESQUISA RENÉ RACHOU, FIOCRUZ
3 INSTITUTE OF BIOLOGICAL SCIENCES, UFMG

Abstract

Trypanosoma cruzi, the etiological agent of Chagas disease is a protozoan that has three developmental forms, which are biochemically and morphologically distinct and programed to rapidly respond to the drastic environmental changes this parasite faces during its life cycle. Unlike other eukaryotes, protein-coding genes in this protozoan are transcribed into polycistronic pre-mRNAs that are processed into mature mRNAs through coupled "trans-splicing" and poly-adenylation reactions. Because of this, control of gene expression relies mainly on post-transcriptional mechanisms that are mediated by RNA binding proteins (RBP) that control steady-state levels and translation rates of mRNAs. We analysed all sequences corresponding to RNA binding motifs by extracting from Pfam database and using these sequences in BLAST searches against all T. cruzi CL Brener proteins. BLAST hits having E values <10-9 and identity = 85% identified 253 sequences in the T. cruzi genome containing RNA recognition motif (RRM), PABP, Alba, Pumillio and Zinc Finger motifs. Using RNA-seq data generated from cDNA libraries constructed with mRNA isolated from epimastigotes, trypomastigotes and amastigotes, we analyzed the expression throughout the T. cruzi life cycle of all sequences containing these RNA binding motifs. Among the genes that are up-regulated in epimastigotes, we identified TcCLB.506739.99, which encodes a RBP containing a zinc finger motif, named TcRBP99. A role of this protein related to parasite differentiation was revealed by the characterization of epimastigotes in which this gene was knocked-out: compared to wild type (WT) epimastigotes, TcRBP99 null mutant showed growth inhibition and reduced capacity to differentiate into metacyclic trypomastigotes. RNA-seq analyses comparing total gene expression of wild type epimastigotes and epimastigotes from two knockout cell lines were performed using a workflow that included mapping of reads to a reference genome using STAR, TopHat2 and Bowtie2 tools and differential gene expression (DGE) analyses were performed with Edge R, limma and Deseq2 packages, having padj < 0.05 and log2FoldChange > 1 as cut-off. Our results revealed 12 genes that showed reduced expression in TcRBP99 knockout cell lines compared to WT. One of them encodes a protein annotated as protein associated with differentiation, whose mRNA is up-regulated in wild type epimastigotes compared to other stages. Immunoprecipitation assays showed that TcRBP99 binds to this mRNA, further suggesting a role of TcRBP99 in controlling the expression of proteins that participate in the epimastigote-trypomastigote differentiation.

Funding: CNPq, FAPEMIG, INCTV

Ab initio prediction of pri-miRNAs based on structural and sequence motifs

Renato Cordeiro Ferreira[1], Alan Durham[1]

1 IME

Abstract

MicroRNAs (miRNAs) are a category of small non-coding RNAs that help to regulate the translation process within the cell. They are originated from a long type of transcript called primary miRNAs (pri-miRNAs), which present a distinctive hairpin loop secondary structure and have a set of conserved motifs. Different proteins use these characteristics to distinguish pri-miRNAs from other similar molecules, so that they can generate the mature miRNAs from them. The aim of this project is to explore these patterns to create an ab initio pri-miRNA predictor. The first step to achieve this goal was to create a simple proof-of-concept classifier that used regular expressions and sequence alignment to select candidate pri-miRNAs. The program was tested on 467,100 segments of size 200 nucleotides (obtained with a sliding window of 100 nucleotides) from the human chromosome 21. It filtered a total of 29 sequences that matched the profile, 6 of which presented high similarity (alignment with e-value less than $10e^{-5}$ against the miRBase database) with sequences annotated in other human chromosomes. This result shows the potential of using these signals to identify likely candidates. The next step will be to implement a full probabilistic model, such as a Context-Sensitive Hidden Markov Model (csHMM), to identify pri-miRNAs. A csHMM will be able to describe the long-range dependencies between positions in the hairpin loop, besides encoding the distribution of nucleotides obtained from real training examples. This way, we expect to create an automatic way to find candidate miRNAs that have not been experimentally observed yet.

Funding: CAPES

7 | Software Development and Databases

EntropyClusterGenes: a R package for clustering genes according ontologies and pathways

André Luiz Molan[1], Carlos Biagi Jr[2], Giordano Bruno[1], Jose Rybarczyk-filho[1]

1 UNESP
2 UNESP - BOTUCATU/SP

Abstract

NGS technologies have transformed the way we study living organisms. By the application of different techniques, such as RNA-seq, numerous species can be studied at a relatively low cost. The amount of data generated, however, is huge. It's not so easy to analyze it, demanding computational tools increasingly efficient and with different approaches, highlighting those with a focus on functional analysis. In this way, we developed EntropyClusterGenes, a R package capable of clustering genes according to their respective Gene Ontology (biological processes - BP, molecular functions - MF, cellular components - CC and KEGG pathways) and determining the significance of such sets based on the expression values ??of their genes. We start with a text file containing a gene list and their expression values in two comparative samples (e.g., experiment and control). Through the clusterProfiler and topGO R packages, linking them to online databases of different species, we group the genes according to their respective functions. The behavior of the genes within each group is shown by the relative calculation of two variables: gene activity and gene diversity. The first one characterizes the set just according to the expressed value of the genes contained therein, while the second one measures the gene diversity within the set by the use of a normalized Shannon's entropy function. Each set is assigned with a p-value, obtained through the bootstrapping statistical method and multiple comparisons between the activity and gene diversity variables. To determine which sets are significant, we apply the FDR (False Discovery Rate) statistical method, considering, by default, values ??lower than 0.05. To demonstrate the use of the EntropyClusterGenes, we applied the tool in two data sets of microarray (Homo sapiens and Rattus norvegicus) and a RNA-seq data set of Aedes aegypti. The number of significant groups found varied according to the species, sample and function studied. We considered a FDR of 0.05 for all species. Despite the number of bootstrapping steps, we used 10,000 for microarray and 500,000 for RNA-seq. In the case of H. sapiens, from a group of 20502 genes, obtained by comparing different doses of etoposide in the liver, for BP, CC, MF and KEGG, were identified, respectively, 5825, 696, 10082 e 305 groups, with an significance average percentage per Gene Ontoloy of 8.52%.

Funding: CNPq processes 473789/2013-2 and 134469/2016-0.

CeTICSdb Database resources and functionalities for the integration of -omics data and mathematical models of signaling networks

Milton Y. Nishiyama-jr[1], Marcelo S. Reis[1], Bruno Ferreira de Souza[2], Henrique Cursino Vieira[3], Daniel F. Silva[4], Inácio L.m. Junqueira-de-azevedo[5], Julia P.c. da Cunha[1], Junior Barrera[6], Leo K. Iwai[7], Solange M.t. Serrano[8], Hugo A. Armelin[1]

1 INSTITUTO BUTANTAN
2 ECC-CETICS, INSTITUTO BUTANTAN
3 LECC-CETICS, INSTITUTO BUTANTAN
4 ESCOLA POLITÉCNICA, USP SÃO PAULO
5 LETA-CETICS, INSTITUTO BUTANTAN, SÃO PAULO, BRAZIL
6 INSTITUTO DE MATEMÁTICA E ESTATÍSTICA, USP
7 LETA-CETICS, INSTITUTO BUTANTAN
8 LETA-CETICS, INSTITUTO BUTANTAN, SÃO PAULO

Abstract

The understanding of biological systems and signaling networks processes constitutes not only a conceptual challenge but a multi-factorial problem if based on different experimental conditions, treatments, time points, etc. The Center of Toxins, Immune-response and Cell Signaling (CeTICS) aims to understand the behavior of biological systems in specific treatments and conditions, using the -omics data and signaling networks analysis; The studies and research in CeTICS project are intrinsically interdisciplinary, which is coupled to the -omics data and heterogeneous knowledge and implies a necessity of data organization and integration to carry out scientific investigations for the generation of new insights and meaningful results. The CeTICSdb aims to provide a dynamic, user-friendly integrated system, for fully support research management, data management, perform customized on the fly analysis, simulations and apply pattern recognition methods for integration of multiple -omics data.CeTICSdb is the core of ARTISiN, an amalgam of repositories and tools, both public and in-house built ones for analysis of signaling networks. ARTISiN will allow a communication between CeTICSdb and SigNetSim, a tool for generation of dynamical models. Moreover, it has been designed for the integration of multi-omic data and mathematical modeling of signaling network. The CeTICSdb has been built with Django (Python web framework) and is composed of multiple components, which will allow to efficiently evolving it into a data management framework, requiring fewer manual changes, especially in the development of new applications. The platform will integrate the data between multiple platforms such as Galaxy and GBrowse, and public components such as Biomodels database, Cytoscape plugin and Mascot. To evaluate and test the platform, we integrated transcriptome and protein expression profiles with Metabolic Pathways to: i) estimate the pathways relative abundance between different conditions; ii) define and compare the functional activity for the pathways in each condition; iii) infer networks based on STRING information. Finally, our mid-term objective is to make the CeTICSdb platform available as a dry lab to the scientific community and the core for ARTISiN. It has already been

Bioinformatics investigation of non-coding RNAs and transposable elements in plants

Daniel Longhi Fernandes Pedro[1], Nicolas Gil de Souza Aoki[1], Alan Péricles Rodrigues Lorenzetti[2], Douglas Silva Domingues[1], Alexandre R. Paschoal[1]

1 UTFPR - PPGBIOINFO
2 USP

Abstract

Non-coding RNAs (ncRNAs) are transcripts that do not encode proteins. There are several classes of ncRNAs, which the most studied are microRNAs (miRNAs). Transposable Elements (TEs) are the major genomic component in eukaryotic genomes. They can comprise more than 45% of human and animal genomes, and in plants, they comprise up to 90% of the genome. Our research group recently developed the PlanTE-MIR DB, the first public database that studies the relationship between miRNA and TEs in plants. In this repository, users can search, extract and analyze these overlapping features in 10 plant species. Now, we intend to evaluate TEs relationship with all ncRNA classes, generating a new version of PlanTE-MIR DB. New bioinformatics analyses will use public genomic data available at Ensembl Plants portal and results will be accessible on a user-friendly website. Three steps cover the workflow of this investigation: a) Curate and intersect ncRNAs and repetitive DNA features from existing Ensembl annotation; b) Perform de novo TE prediction in plant genomes and intersect ncRNA annotation in order to find new potential overlaps; and c) Compare newly discovered TEs against public ncRNA databases. From 44 genomes available at Ensembl Plants, 25 species have ncRNA annotation. In 24 we found overlap with TEs. The species with most overlapped regions was Zea mays with 3,105 hits and the species with less hits was Sorghum bicolor, with one hit. Finally, we intend to develop a new method to identify plant TEs using deep learning techniques. These computational analyses will provide to the scientific community a friendly way to work with this knowledge.

Funding: UTFPR

A shiny app for the integration and enrichment analysis of genomic region sets by NGS data

Davi Toshio[1], Henrique Cursino Vieira[2], Christiane Bezerra de Araujo[3], Maria C. Elias[3], Bruno Ferreira de Souza[4], Hugo A. Armelin[1], Milton Yutaka Nishiyama Junior[5]

1 INSTITUTO BUTANTAN
2 LECC-CETICS, INSTITUTO BUTANTAN
3 LECC-CETICS, BUTANTAN INSTITUTE
4 ECC-CETICS, INSTITUTO BUTANTAN
5 LETA-CETICS, INSTITUTO BUTANTAN

Abstract

The new biotechnology advances has allowed studies of the systems involved in the DNA integrity, stability, replication, demethylation; recent discoveries have related them to genomic and transcription stability, besides relations between inflammation and DNA damage, which are essential aspects of molecular biology that underlies developmental processes and disease etiology. This work is part of The Center of Toxins, Immune-response and Cell Signaling (CeTICS), which aims to understand the behavior of biological systems based on analysis of -omics data and signaling networks. Several genomic techniques including MFA-seq, MNase-seq, ChIP-seq, DNase-seq and ATAC-seq have been developed to experimentally identify genome-wide profiles of regulatory regions and experiments have been profiled by the CeTICS project and thousands of samples can be found in the ENCODE and Roadmap Epigenomics consortia. The genomic datasets has grown rapidly and has been used as reference databases; they are essential to retrieve information on gene name, protein product, transposable elements, motifs, molecular markers and others and are important in the approaches to identify enriched regions. To allow the identification and characterization of enriched regions from high-throughput sequencing data, which can be increased with visual inspection of the data, we present a Shiny application for an interactive representation and analysis tools based on Fold Change and MACS2 software peak predictions. The app has been developed in R language, using the Shiny framework, integrating multiple Bioconductor packages. The first step is the alignment and coverage estimation, followed by the upload to the results and annotation file to the app for the downstream statistical and graphical analysis. The input files are composed by: i) the fold change and coverage estimation; ii) the MACS2 peak detection and coverage estimation; ii) Genome size and annotation. As a study case, was used the investigation of the DNA replication features of T. cruzi in order to verify possible links between genetic instability and DNA replication, using the high throughput analysis approach MFA-seq (Multiple Frequency Analysis). To do that, epimastigotes of T. cruzi were sorted in early S and G2/M phases and the DNA was extracted from each group and analyzed by MFAseq. The aim of this integrative approach is to allow the identification of enriched genomic regions distinct to each pair of conditions, integrating multiple annotations, coverage, predicted peaks, SNP's, genes, transcripts and others and the integration of Machine Learning methods to improve the data integration analysis.

Funding: #2013/07467-1, São Paulo Research Foundation (FAPESP)

Crowdnotation: A Crowdsourcing Annotation Tool for Genomics Studies

Diogo Matos da Silva[1], Helder Takashi Imoto Nakaya[1]

1 USP

Abstract

When authors deposit a microarray study into GEO database, they are free to describe the experiments using their own words. This author-based annotation method follows no defined ontology or classification standards, creating enormously difficulty for others when querying studies based on specific key words. Only human manual curation can properly annotate studies, as several computer programs have already attempted and failed at this task. We are developing a web-based tool named Crowdnotation, where students can remotely annotate the studies. In return, any student participating in the annotation will be included as an author in our publications. We believe that this alone serves as a strong incentive for a significant number of students. More importantly, to make the annotation process more attractive and enjoyable, Crowdnotation will have several gamification features, such as scores, rankings, friends and badges. This will promote engagement and ultimately improves performance and efficiency. Annotation accuracy will be ensured through consistency among peers. Students will learn about the different types of experimental designs and will gain significant knowledge in the field of modern molecular biology. We expect this worldwide web-based community strategy will create a massive collaborative network of students working towards a common goal. In addition, students from remote areas of developing countries will gain a valuable opportunity to be part of a scientific collaboration and subsequent publication. In the future, this tool may be easily adapted to answer various scientific questions and for that, we will have the contact information of several thousand future scientists. Finally, all annotation data will be organized into a structured and open-accessed database, and the codes for Crowdnotation will be freely available for future developers.

Funding: University of Sao Paulo

Classifying gene mutations in the scientific literature using neural network

Douglas Teodoro[1], Luc Mottin[2], Anaïs Mottaz[2], Paul Van Rijen[2], Emilie Pasche[2], Julien Gobeill[2], Patrick Ruch[2]

1 SIB SWISS INSTITUTE OF BIOINFORMATICS
2 HES-SO/HEG GENEVA

Abstract

Discriminating between mutations that contribute to tumor growth and neutral mutations is essential for the success of precision medicine. Currently, the interpretation of genetic mutation is done by clinical pathologists via manual reviews of the scientific literature. In the context of the Classifying Clinically Actionable Genetic Mutations competition track, we investigate machine learning methods to automatically classify nine categories of genetic mutations present in text-based clinical literature. Given a scientific text article and a gene-variation pair, described in the article, our algorithm predicts the probability that the article provides evidence for the nine mutation classes. We use the paragraph2vec algorithm to embed the text in a vector space and use the vectors as features for the machine learning algorithm. The articles are divided into three parts: containing evidence to relevant gene-variation mutation pair, containing evidence to non-relevant gene-variation mutation pair, and not-containing evidence to gene-variation pair. To train and assess the methods, we use an expert-annotated dataset containing 3321 variant annotations provided by Memorial Sloan Kettering Cancer Center. We compare neural-based methods, such as Multi-Layer Perceptron (MLP) and Convolution Neural Networks, and tree-based methods, such as Random Forest and Extreme Gradient Boosting, against a Naïve Bayes baseline. Our best method (MLP) achieved an average precision of 0.7101 (0.9658 multi log-loss) compared to the 0.6220 average precision (1.1870 multi-log loss) of the baseline method. We are working to improve the classification errors by bringing further domain knowledge into the classifier. We expect that such methods could be useful for identifying relevant articles for manual curation.

Funding: ELIXIR-EXCELERATE/676559

R package development to analyze the cancer genome atlas data: a study case based on hypoxia induced factor- a3 isoforms

Fábio Malta de Sá Patroni[1], Douglas Adamoski[2], Marcelo Falsarella Carazzolle[3], Sandra Martha Gomes Dias[4]

1 GRADUATE PROGRAM IN GENETICS AND MOLECULAR BIOLOGY, INSTITUTE OF BIOLOGY UNICAMP, CAMPINAS
2 GRADUATE PROGRAM IN GENETICS AND MOLECULAR BIOLOGY, INSTITUTE OF BIOLOGY UNICAMP
3 BIOLOGY INSTITUTE - UNICAMP, NATIONAL CENTER FOR HIGH PERFORMANCE COMPUTING/UNICAMP
4 BRAZILIAN NATIONAL CENTER FOR RESEARCH IN ENERGY AND MATERIALS, BRAZILIAN BIOSCIENCES NATIONAL LABORATORY

Abstract

A great magnitude of information on multi-resource omics data is being created and made freely available through the project The Cancer Genome Atlas (TCGA). Although the amount of data rises every year, data mining tools are not following at the same pace. The efficient use of this information, also called the "big data", has the potential to unveil new observations and mechanisms that can impact on cancer treatment. TCGA data are stored at the GDC Data Portal and GDC Legacy Archive, both of which hosted by the US Nacional Cancer Institute (NCI). The totality of the data comprehends genomic, transcriptomic, proteomic and metilome information from 33 types of cancer. Given the complexity and extension of the available data, new analytical tools are necessary to automate and facilitate the data mining process. R programming language is being widely used for dealing with "big data". This work has two main goals: First, to create an R package aiming at to download, organize, analyze and report TCGA data; second, apply the package to identify potential downstream targets of the transcriptional factor HIF-3a isoforms. HIF regulates the expression of genes as a response to hypoxia and is an important player on the tumor metabolism adaptation process. Our R package, GDCRtools (version: 0.0.9) was developed and used to analyze the HIF3a2 isoform in four tumor types: Ovarian serous cystadenocarcinoma [OV], Testicular Germ Cell Tumors [TGCT], Uterine Carcinosarcoma [UCS] and Stomach adenocarcinoma [STAD].Tumors were divided among higher and lower HIF3a2 expression, and differential gene expression determined. Gene Ontology and Reactome was employed for pathway enrichment analysis and revealed enriched terms related with extracellular matrix organization, blood vessel development and GPCR downstream signaling, potentially linking this isoform with these processes.

Funding: This work was supported by grants from São Paulo Research Foundation (2015/26059-7)

BioFeatureFinder (BFF): Flexible, unbiased analysis of biological characteristics

Felipe Ciamponi[1], Michael Lovci[1], Katlin Massirer[1]

1 CBMEG - UNICAMP

Abstract

BFF interrogates interesting genomic landmarks (ex. alternatively spliced exons, DNA/RNA-binding protein binding sites, and gene/transcript functional elements) to identify distinguishing biological features (nucleotide content, conservation, k-mers, secondary structure, protein binding sites and others). BFF uses a flexible underlying model, combining classical statistical tests with big data machine learning strategies, that takes thousands of biological characteristics (features) and can interpret category labels in genomic ranges or numerical scales from genome graphs. The algorithm is python-based with scalable multi-thread capabilities, designed to be compatible with a wide array of servers ranging from notebooks to HPC clusters. Due to flexible nature of it's design, BFF can also be easily modified to include new functions and sources of data in it's analysis process. As proof-of-concept, we applied BFF to an eCLIP-seq (enhanced crosslinking-immunoprecipitation followed by RNA-seq) dataset for the mRNA targets of RNA-binding proteins (RPBs) RBFOX2. Our algorithm was capable of recovering several major features described previously in the literature, as the GCAUG binding motif for the RBFOX2 protein. To showcase the potential uses of BFF, we analyzed 112 eCLIP-seq datasets from RBPs available at ENCODE, identifying biological features associated with the binding sites for these proteins. From a total of 5498 input features, BFF predicts an average of 624 important features for each RBP. 98 RBP binding maps that were marked by co-location with other RBP binding maps, with known complexes (ex. IGF2BP1-3, U2AF1-2) being identified by BFF. 40 RBP binding maps were marked by their relative abundance of sequence motifs, with known examples from the literature (ex. TARDBP, SRSF1, PUM2, PTB and QKI) being successfully recovered by BFF. Secondary RNA structure was a distinguishing feature for 64 proteins, some which are known RNA-structure binding proteins (ex. TAF15, KHDRBS1 and ESWR1). Taken together, our results show that BFF provides a flexible and reliable analysis platform for large-scale datasets, while at the same time providing a way to control observer bias and uncover latent relationships in biological datasets.

Funding: FAPESP, CNPq

Integration and Data Mining in Drug Target Detecting for Schistossoma mansoni

Francimary Procopio Garcia[1], Kele Teixeira Belloze[1]

1 CEFET/RJ

Abstract

Classified as a neglected disease, in spite of it's acting in underdeveloped countries, schistosomiasis, caused by Schistossoma Mansoni, is considered one of the most important endemic diseases in the world, having an estimated number of around 240 million infected and 700 million people living in an area with a high risk of transmission. There's currently one sole drug recommended by the World Health Organization for schistosomiasis treatment which, although being effective in the elimination of the vermin, it shows collateral effects and can only act upon its mature form. Therefore, researching for new alternative drug targets in combating schistosomiasis is required. This work has as main objective the identification and classification of possible new drug targets for S. mansoni. The proposed methodology for the development of this work is described as follows. At first the identification of ortholog proteins between S. mansoni and three eukaryotic models organisms will be done: Caenorhabditis elegans (nematode), Saccharomyces cerevisiae (yeast) and Mus musculus (mouse), based on the concept of gene essentiality. Subsequently, the process of identifying homologous proteins between the S. mansoni proteins raised in the previous step and druggable proteins (targets for drugs), available at Drugbank and Therapeutic Target Database (TTD) databases, will be conducted. These two steps will result in an intermediate database composed of essential and druggable candidate proteins of the organism under study, represented by primary sequences integrated with the aggregate annotations during the accomplishment of these two steps of the methodology. From the candidate proteins raised, the research will proceed on identifying information of these protein's secondary structures, in order to enrich the database conceived. For this step a homology based approach will be adopted using the secondary protein structures available at Protein Data Bank (PDB). Data from this integrated database will be categorized using frequent patterns models such as Apriori to identify consistent behaviors among candidate proteins and provide them with an druggability index. A decision tree model will be used to identify the candidates with the highest combination weight and validated through cross validation functions, where the available data will be divided into two mutually exclusive subsets, one for training (parameter estimation) and another for testing (validation). The percentage of candidates prediction with the highest druggability will be calculated and their druggability indexes will be validated and discussed according to data obtained in the literature. As a result of this work, it is expected to obtain a list of S. mansoni proteins which may be indicated as drug targets, and thus contribute to the initial step of the drug development process. This work is in an initial phase of studies in which a literature review is being carried out.

Funding: CEFET/RJ

Data integration of Pseudomonas aeruginosa CCBH4851 genome sequence to support a whole cell modelling

Ribamar Santos Ferreira Matias[1], Francimary Procopio Garcia[1], Kele Teixeira Belloze[1]

1 CEFET/RJ

Abstract

Pseudomonas aeruginosa is a bacterium species that arouses great interest, both in scientific and public health agencies, due to its strong association with pathogens related to hospital infections. A strain of this species, Pseudomonas aeruginosa CCBH4851, was found in Brazil in 2008, and when tested, was resistant to several antibiotics, of which only one, polymyxin B, was able to combat it effectively. Studies on this bacterium, aiming the construction of its whole-cell model, are being conducted by researchers of the Oswaldo Cruz Foundation. Such studies are intended to better understand the behavior of bacteria and thus make suggestions for new drug targets. The objective of this work is to integrate genomic sequencing data of the bacterium Pseudomonas aeruginosa CCBH4851 with data from Pseudomonas aeruginosa PAO1, a reference bacterium in the study of Pseudomonas sp and Escherichia coli, a bacterium that is a model organism. Based on the data integrated, it will be developed a knowledge base to support the identification of regulatory and metabolic pathways of the complete cell model of this bacterium.The proposed integration will be based on Gene Ontology Consortium's Database informations, known as GO Database, which is a public data repository, composed of ontologies and gene annotations in terms of these ontologies. The proposed methodology for constructing the integrated knowledge base will be divided into the following actions: i) extract, transform and cleaning data sequences and annotations of P. aeruginosa CCBH4851; ii) compare the sequences with the E. coli and P. aeruginosa PAO1 models; iii) annotate the compared sequences using Gene Ontology Database; iv) associate the results with the discovered data of the regulatory and metabolic pathways. v) associate and validate data in the literature. Based on the result acquired, a knowledge base will be developed in order to facilitate the research results, presenting the creation of an uniform information set, with terms widely accepted and known by the scientific community. In addition to the availability of this knowledge base, it is expected that the integrated information of the study and reference bacterium will support decision making in the assemblages of the regulatory and metabolic pathways of P. aeruginosa. This work is in its initial development stage in which the first methodology step is being carried out as well as the literature review.

Funding: CEFET/RJ

Active Semi-Supervised Learning for Analysis of Biological Data

Guilherme Camargo[1], Pedro Henrique Bugatti[2], Priscila T M Saito[2]

1 PROGRAMA DE PÓS-GRADUAÇÃO EM BIOINFORMÁTICA - PPGBIOINFO, UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ, CORNÉLIO PROCÓPIO
2 FEDERAL UNIVERSITY OF TECHNOLOGY - PARANA

Abstract

In the last few years, new data capture devices have made it possible a major technological breakthrough. Thus, large complex databases (e.g. images, sounds or texts) are obtained daily. In order to allow the storage and the retrieval of information from these databases, it is necessary specialists to annotate the samples. However, the annotation by specialists can bring inconsistencies to the samples, since different individuals can interpret the samples in divergent ways. Another reason to consider is the cost to perform the task of annotating the samples, which is high and exhaustive to specialists. Therefore, a solution to the problem would be to automate the process of identifying the most informative samples using computational methods. In this way, a label is assigned to each sample, classifying it according to the scope of the problem. One way to develop a suitable solution is applying machine learning techniques in order to build a pattern classifier. To take benefit from the large number of unsupervised samples available in disproportion to the scarcity of supervised ones, semi-supervised learning techniques have been explored using partially supervised and unsupervised training samples, where supervised samples propagate their labels to the unsupervised ones. However, such techniques neglect the existence of redundant samples, as well as the existence of more relevant samples that could boost the classifier learning. In this context, active learning techniques associated with semi-supervised techniques are interesting, since a smaller number of more informative samples automatically selected through the active learning strategy, and then annotated by a specialist can propagate the labels to a set of unsupervised samples (through the semi-supervised learning strategy). Therefore, we developed a new active semi-supervised learning approach for biological data, exploring new strategies for selecting more informative samples for the classifier learning. Preliminary results show that the union of active and semi-supervised learning improves accuracy for some biological datasets, reaching higher values faster, showing that the obtained active semi-supervised classifiers are more efficient than the supervised and the semi-supervised ones.

Funding: CNPq (#431668/2016-7, #422811/2016-5), CAPES, Fundação Araucária, SETI, and UTFPR.

Identification and Visualization of Expression Patterns by the Integration of Pathways, Transcriptome and Proteome profiles

Henrique Cursino Vieira[1], Bruno Ferreira de Souza[2], Hugo A. Armelin[3], Milton Yutaka Nishiyama Junior[4]

1 LECC-CETICS, INSTITUTO BUTANTAN
2 ECC-CETICS, INSTITUTO BUTANTAN
3 INSTITUTO BUTANTAN
4 LETA-CETICS, INSTITUTO BUTANTAN

Abstract

The integration of multi-omics data is a great challenge and is necessary for the understanding of the complexity of biological systems and signaling network interactions. Discovering the gene or protein expression signature associated with a specific treatment or condition is the basic question. However, the studies based in only on molecular level (e.g. genome, transcriptome, or proteome) may be incomplete and fail to reveal the multi-layer interactions in the different molecular levels. The computational approaches for good predictions and efficient data integration are not well established yet and can be dependent on the specific experimental design. Many scientists are looking for integrated and graphical tools for visualization of networks, pathways and generate pictures and tables presenting their findings. This work is part of The Center of Toxins, Immune-response and Cell Signaling (CeTICS), which aims to understand the behavior of biological systems based on analysis of -omics data and signaling networks; The studies and research in CeTICS are intrinsically interdisciplinary, which is coupled to the -omics data from genomics, transcriptomics and proteomics, heterogeneous knowledge. The necessities to the integrative analysis of such multiple layers, is especially focused on the most preferred methods for understanding global gene regulation: high-throughput RNA sequencing and mass spectrometry (MS) expression profiles, coupled to the integration with the relative abundance of metabolic and signaling pathways, which can be increased with the application of machine learning methods and viewers for the inspection and analysis on-the-fly. Data visualization tools can be categorized into two types, although they can overlap: tools focused on automated methods for the interpretation and exploitation of large biological networks; and tools for assembly and validation of tracks. Thus, we present a Shiny application for an interactive viewer with automated analysis tools, which will enable novel conclusions to be drawn from transcriptomic and proteomic integrative analysis. The researcher will can interact with the application, so it can view the content in a clean and fluid way, making possible aggregate data from different biological sources. The goal is beyond visualization, and will allow performing multivariate analysis for systems-level, understanding from multi-dimensional data and be optimized to show the results through the graphs. Finally, our mid-term objective is to develop an integrative approach, which will be aggregated to the CeTICSdb platform, which will be available to the scientific community.

Funding: #2013/07467-1, São Paulo Research Foundation (FAPESP)

Deep Learning Strategies for Autism Severity Classification in Children

Hudson Pereira[1], Priscila T M Saito[1], Pedro Henrique Bugatti[1]

1 FEDERAL UNIVERSITY OF TECHNOLOGY

Abstract

Autism Spectrum Disorder (ASD) is a syndrome characterized by difficulties in social interaction, qualitative deviations in communication and repetitive behaviors. This syndrome is also defined as loss of contact to reality, caused by impossibility or great difficulty in interpersonal communication. ASD is classified into three degrees of severity: mild, moderate and severe. The early diagnosis of the child with autism is essential for an effective treatment. In children under three years, it is possible to achieve an improvement of 80%. In children up to five years, an improvement of 70% can be obtained, and above that age, any treatment is compromised. Literature studies generally consider several techniques for diagnosis. However, they do not take into account the identification of the severity degrees, as well as the differences between boys and girls with ASD. Therefore, this work aims to develop a computational method to diagnose and classify the autism severity degrees. Moreover, it is intended to propose strategies in order to identify possible differences in facial micro-expression between boys and girls, since the diagnosis in girls is more difficult. The methodology to be developed consists of: (I) obtaining images with frontal pose of children between 3 and 5 years; (ii) extracting micro expressions through the Histogram of Oriented Optical Flow (HOOF) algorithm; (iii) extracting facial expressions using Convolutional Neural Networks (CNNs); (iv) identifying autism severity degrees and differentiating boys and girls with autism using different classifiers, such as Support Vector Machines (SVM).

Funding: CNPq, (#431668/2016-7, #422811/2016-5), CAPES, Fundação Araucária, SETI, UTFPR, and PPGBIOINFO

CINDEX: a software for protein ranking through network modeling based on graph theory

James Shiniti Nagai[1], Hugo Rody Vianna Silva[1], Alexandre Hild Aono[1], Estela Araujo Costa[1], Reginaldo Massanobu Kuroshu[1]

1 UNIFESP

Abstract

Metabolic networks have increased in complexity throughout the evolution of species becoming strongly connected metabolic blocks, which seems to have given stability to the flux of information in these networks and, possibly, turning organisms into more relaxed ones to adapt. We present the first version of CINDEX, a software for protein ranking through the modeling of protein-protein interaction (PPI) networks based on graph theory; it can be downloaded at https://github.com/hugorody/cindex. The tool accepts as input an organism's subset of proteins provided by the user and uses PPI information from the KEGG Pathway database to model a specific metabolic network throughout a directed graph. The proteins are set as the nodes of the graph, whereas the connections among the nodes are given by arcs (directed edges) that are created when the product of a protein is a substrate to another. CINDEX then calculates the centrality degree of nodes using three different metrics - Degree Centrality, Closeness Centrality, and Betweenness Centrality -, which provides to the user different biological perspectives for protein (node) ranking. Additionally, our software searches for lethal bottleneck proteins - proteins represented by nodes that disconnect the network when removed, thus essential to keep the flux of information in a network and whose inactivation could be lethal to the organism. Finally, the clustering coefficient is calculated to indicate the presence of protein clusters within the networks; a subset of interacting proteins likely to control many cellular processes.

Funding: CAPES

A novel noninvasive prenatal paternity test using microhaplotypes

Jaqueline Yu Ting Wang[1], Anatoly Yambartsev[2], Renato Puga[3], Martin R. Whittle[4], André Fujita[2], Helder Takashi Imoto Nakaya[1]

1 USP
2 INSTITUTE OF MATHEMATICS AND STATISTICS - USP
3 HOSPITAL ISRAELITA ALBERT EINSTEIN
4 GENOMIC ENGENHARIA MOLECULAR

Abstract

Paternity tests are usually done by analyzing DNA samples from the alleged father, the mother, and the child. To perform this exam before the birth, invasive methods such as amniocentesis and chorionic villus sampling are usually used. Fortunately, the discovery of fetal DNA (fetal cell-free DNA, fcfDNA) in maternal plasma and serum, and the development of techniques to analyze this fcfDNA have allowed researchers to reduce this risk for both fetus and mother. Although paternity tests that analyze Short Tandem Repeats (STRs) from fcfDNA are possible, they are not reliable because DNA degradation often occurs. SNPs (Single Nucleotide Polymorphisms) have been demonstrated to be good candidates for human identification and they can be obtained from small DNA fragments (even from degraded DNA). To increase the number of possible genotypes and decrease the amount of analyzed SNPs, our analyzes focus on microhaplotypes. Microhaplotypes are chromosomal segments smaller than 200 base pairs (bp) containing two or more SNPs that form at least three distinct haplotypes. Since fcfDNA has approximately 145 bp, this is sufficient to contain microhaplotypes that can be sequenced using Next Generation Sequencing (NGS) technology. The aim of this project is to determine the probability of paternity using SNPs within microhaplotypes. Microhaplotypes were chosen based on previous literature review. The haplotypes frequencies were calculated based on the ethnic groups from 1000 Genomes database. To accomplish this objective, raw DNA sequence data from three DNA samples were analyzed: the alleged father, the mother, and the maternal plasma (mixture of mother and fetus cell-free DNA). Then, using a script developed based on SAMtools and Perl programming language, we obtained the genotypes of the alleged father and mother, for each microhaplotype. By combining genotypic information, population frequencies, and fetal fractions (plasma), we developed a method to calculate the probability of paternity in cases of non-exclusion.

Funding: Genomic Engenharia Molecular

Updated TAXI, a taxonomic innovations database depicting operons structure and evolution

Lucas Ferreira[1], José Miguel Ortega[2]

1 SGC - STRUCTURAL GENOMICS CONSORTIUM, UNICAMP
2 UFMG. LABORATÓRIO DE BIODADOS

Abstract

The structure of operons, navigation from one operon to another that shares orthologous genes and the analysis of the clade of origin of the operon and their component genes comprise the information present in TAXI database. TAXI stands for Taxonomic Innovations, being the main goal to understand along evolution how the operons have formed, which one is the most recent and most ancient gene in it, and understanding the funcions that are restricted to anym microbial clade, e.g. a family, a species or even a strain. Update of TAXI database presents a new set of fully indexed tables. The present number of organisms, transcription units, clusters of orthologous genes and genes add up to, respectively: 1753, 3343458, 551692, and 6732117. The most important updates refer to integration of identifiers with external databases UniProt and Kegg. TAXI information can be accessed through an organism of interest, gene symbol, UniProt accession or Kegg Orthologues group (KO). Queries to the database return operons where the most recent or the most ancient gene is in the first position, aiming to facilitate the study of the evolution of regulation in operons. By using the navigation though orthologue groups it is possible to verify the distinct operon compositions of the gene of interest, helping the study of co-functionalities. TAXI database is now available in an instance located at a CPD at bioinfo.icb.ufmg.br/taxi.

Funding: FAPEMIG

RTranscriptogram: a tool for biological data integration

Alex Augusto Biazotti[1], Túlio Moreira Fernandes[1], André Luiz Molan[2], Agnes Alessandra Sekijima Takeda[1], Jose Rybarczyk-filho[2]

1 INSTITUTO DE BIOCIÊNCIAS DE BOTUCATU - UNESP
2 UNESP

Abstract

Every day, new technologies are emerging that make it possible the large-scale study of RNAs transcribed by an organism under specific conditions, providing a huge amount of information. However, the traditional methodologies are not able to efficiently analyze these data due the use of pre-defined cut-offs, thus eliminating a large number of genes not considered differentially expressed, and consequently reducing precision and accuracy of the study. This work proposes the tool called RTranscriptogram which performs an overall analysis of an organism, integrating protein networks biological, processes and expression genes.. This tool clusterize the network, extract the group and their respective biological information and project the expression genes on the network. To test the tool, we used the STRING database version 10 to prospect for a protein network for Homo sapiens, Gene Ontology provided the biological processes (BPs) for this study. Gene expression data were prospected from Gene Expression Omnibus (GEO): GSE19804 - lung samples from Taiwanese female nonsmokers with and without cancer. GSE10072: - lung samples from Italian female and male smokers, former-smokers, non-smokers with or without cancer. These databases were integrated by the transcriptograma technique to obtain expression profiles correlated with protein networks and ontologies. The analyzes presented 2 up-regulated biological processes and 30 down-regulated biological processes that were similar in the comparisons made with people with cancer, in addition the transcriptional activity for Taiwanese and Italians presented a similar profile. Smokers with cancer presented 175 altered BPs when compared with non-smokers. Despite different habits among populations, lung cancer has a high similarity in transcriptional activity.

Funding: CNPq processes 473789/2013-2

Heart Rate and its Variability as Predictors of Activities and Controls for Simple HMI

Juliana Cavalcanti[1], Andre Fujita[2]

1 USP
2 IME - USP

Abstract

Several human-machine interfaces have been devised making use of the EEG signal. However, because of its distance to the brain, which is surrounded by the skull, this signal is susceptible to a high ratio of noise, and also has the disadvantage of requiring uncomfortable electrodes to be attached to the subject. Electrical signals from the heart are less noisy, obtainable by more comfortable chest straps and easily processed in order to extract the interval between peaks. Such intervals offer not only information on instant heart rate, but also its variability, which can then be used to detect subtle variations caused by the autonomous nervous system, responsible for modulating the heart rate and other involuntary body conditions. Our aim is to investigate the feasibility of using data from heart rate variability in predicting the activities executed by an individual and, eventually, controlling a simple Human-Machine Interface, or as a context provider for more complex EEG-driven Brain-Machine Interfaces. In order to achieve that, we will develop a database including heart rate and annotations with a broad description of subject's daily activities, and then analyze the data contained in this pilot database to search for information that helps us predict activities using only the heart rate signal. Furthermore, we will implement a simple game, which will use the heart rate variability as input, serving as a proof of concept to evaluate whether heart rate is fit for the goals described. For this game, subjects will train the ability to voluntarily control their heart rate, and we will analyse the signal provided in this training in search of some mechanism involved that can help us detect the intent to alter heart rate within a reasonable time frame, allowing the responsive control of an interface.

Funding: -

Rational design of profile HMMs for viral detection, classification and discovery

Liliane Santana Oliveira Kashiwabara[1], Dolores U. Mehnert[1], Paolo M. A. Zanotto[1], Alan Durham[1], Alejandro Reyes[1], Arthur Gruber[1]

1 USP

Abstract

Some of the most devastating pandemic diseases have arisen through the transmission of emerging viruses that have not been detected before the tragic consequences of their dissemination. The detection of novel viruses is a challenging task due to the their high evolutionary rates. Profile HMMs are a powerful way of modeling sequence diversity and constitute a very sensitive approach to detect emerging viruses. In this work, we report the development and implementation of TABAJARA, a tool for rational design of profile HMMs. Starting from a multiple sequence alignment (MSA), TABAJARA is able to find blocks that are either (1) conserved across all sequences or (2) discriminative for two specific groups of sequences. For the identification of regions conserved across all protein sequences of an MSA, we implemented a previously described algorithm based on Jensen-Shannon divergence. In the case of nucleotide sequences, TABAJARA can use Shannon entropy or, alternatively, different substitution matrices to define position-specific scores. To find group-discriminative blocks, the program uses Mutual Information or Sequence Harmony for both, DNA or protein sequences. Once position-specific scores have been determined, TABAJARA uses a sliding-window to screen the whole alignment and delimit top-scoring regions. The program automatically extracts the selected alignment blocks, discards identical sequences, and builds the corresponding profile HMMs, which can then be used for many potential applications. To validate such models for viral detection, classification and discovery, we used two different viral taxonomic groups: phages of the Microviridae family and viruses of the Flavivirus genus. Using different metagenomic datasets, we observed that profile HMMs generated by TABAJARA can successfully be used as seeds to reconstruct genome sequences with GenSeed-HMM program. In both viral groups, we were able to obtain wide-range seeds (generic for all members of Microviridae or Flavivirus); and narrow-range seeds, exclusive to specific Microviridae subfamilies (Alpavirinae and Gokushovirinae) or to particular flaviviruses (e.g. DENV, ZIKV or YFV). The approach proposed here, using short and specific sequences to build profile HMMs, represents a radical change, compared to viral models from public databases such as vFam and pVOGs, built from MSAs derived from full-length protein sequences. Narrow-range seeds can be used to detect and classify already known viruses, whereas more generic seeds are useful for detecting wider viral groups, as well as distantly related viruses that could represent potentially emerging pathogens.

Funding: PhD fellowship from CAPES (LSO)

Output Organizer - a software to facilitate POTION results interpretation

Mariana Teixeira Dornelles Parise[1], Doglas Parise[1], Marcus Vinicius Canário Viana[2], Anne Cybelle Pinto Gomide[2], Vasco Ariston de Carvalho Azevedo[1]

1 UFMG

Abstract

Positive selection studies have been used to identify genes involved in the emergence of new phenotypic traits, speciation and host-pathogen interaction. The automatic detection of positive selection can be performed using POTION software, which is an end-to-end pipeline to genome-scale analysis. This software allows users to configure and run the analysis easily and quickly, but gives the results in many files. Although the pipeline gives the most important information in some files, examining the other result files for more details required to write a manuscript is time-consuming and may be error prone due to human errors. In order to facilitate researcher's information retrieval, a software to organize and summarize the most relevant information in the POTION results was created. This solution was developed using Java in the NetBeans IDE. It receives the POTION log and positive.out files, the directory containing the intermediate files of the analysis, a file containing gene homology relations in the OrthoMCl 1.4 format, the folder which contains the nucleotide fasta files and the number of the organism genetic code. Using the given files, the software gives to the user an overview showing eight key features of the analysis and a table showing how many genes were pre-filtered and the reasons why. In addition, the program creates a file for each positive selected group showing the significant statistic model, the Bayes Empirical Bayes results table and a detailed table for each codon position under selection. The presented software facilitates the interpretation of the POTION results, giving to the final user an easy, organized and brief overview of the analysis.

Funding: CAPES, CNPq, FAPEMIG

Decision-making model for the monitoring and identification of risk groups for Type 2 Diabetes Mellitus comorbidities using Fuzzy NN algorithm

Melissa Mello de Carvalho[1], Waldemar Volanski[1], Geraldo Picheth[1]

1 UFPR

Abstract

Type 2 diabetes mellitus (DM2), its comorbidities and complications represent avoidable expenses for patients and public health. Complication prevention can be proposed with a decision model to identify risk groups. The present study aims to study the complications of DM2 from a database with 209 female patients (40 to 87 years) with DM2 at a mean of 12.2 years (1 to 40 years of the diagnosis). The study is approved under the CAAE: 1038112.0.0000.0102. We analyzed 29 attributes arranged in biochemical, anthropometric and monitoring attributes of DM2, such as time of diagnosis in years, presence or absence of comorbidities. The data do not contain missing data, the attributes have multivariate numerical and nominal variability (Y/N). The comorbidities are: coronary artery disease (CAD), retinopathy, neuropathy, nephropathy. They are classified with the Fuzzy NN algorithm from biomarkers. HbA1c = 7% classifies optimal glycemic control in patients with DM2 according to SBD 2015-2016. The choice for fuzzy systems is due to the classification diffusion of attributes with high variability and multicriteria decision. The data were classified with the Fuzzy NN algorithm in the WEKA software, trained and cross validated with the following specifications: using 10 nearest neighbors for classification, Similarity measure 4; Implicator Gödel; T-Norm Algebraic; Relation composition Lukasiewicz. The classes were modified to classify the objects of study: Control, Retinopathy, Neuropathy, Nephropathy and CAD, omitting in each analysis glycemic variables. Fuzzy NN was able to identify all cases of neuropathy and nephropathy as presented in the data with about 93.3% and 94.5% accuracy respectively. With the retinopathy class, it was possible to identify all previously known cases (53 cases) and predict another 5 cases of risk with 72.2% accuracy, 74% specificity. With the Control class, the accuracy was 76% with a precision of 81% and a specificity of 70%. By including the variable insulin use the accuracy increases to 87.5%, precision: 88% and specificity 90%. In all cross validations values the classification coverage of the Fuzzy NN remained above 98%. The classification Control with and without insulin indicates the importance of the medical monitoring in DM2, approximately 32% of the patients besides presenting poor glycemic control do not use insulin nor hypoglycemic agents, which represents a deficit in the follow-up of this portion of patients.

Funding: None

PFstats: An Open Tool for Evolutionary Protein Analysis

Néli José da Fonseca Júnior[1], Marcelo Querino Lima Afonso[2], Lucas Bleicher[1]

1 UFMG

Abstract

PFstats is a software developed for the extraction of useful information from protein multiple sequence alignments. By analyzing positional conservation and residue coevolution networks, the software allows the identification of structurally and functionally important amino acid groups and the discovery of probable functional subclasses. Furthermore, it contain tools for the identification of the possible biological significance of these findings. The goal of this project is to provide a computational tool with interactive graphical user interface and data visualization tools to predict global and specific functional amino acid residues and also find functional subclasses in protein families. The software was developed under a client-server architecture. The client was developed in C++/QT and in the server side a java webservice is made to enable the communication between the client and repositories databases of UniprotKb, PFAM and PDB. PFstats includes methods for alignment filtering, residue conservation and coevolution analysis, automatic UniprotKb queries for residue-position annotation, amino acid alphabets reduction and many possible data visualizations. We have studied four protein family domains: lysozyme C/Alpha-lactoalbumin, phospholipases A2, nitrogen regulatory protein PII, and the DNA binding domain of the nuclear receptors IV. In all of them communities of residues related to catalytic and binding sites were found, and also communities related to structural importance, as hydrophobic putative channel and secondary structures, and communties reated to taxonomic specificity. PFstats is free and open source, being distributed in the terms of the GPLV3 licence. The software is available in GUI and terminal versions at http://www.biocomp.icb.ufmg.br/biocomp/software-and-databases/pfstats/. We provide binaries for Windows and Linux (debian), but also compilation instructions for other systems, in addition to the source code and a manual.

Funding: CAPES and FAPEMIG

Biological data exporting tool

Yoshin Efrain Contreras Oscco[1], Giovana Secretti Vendruscolo[1], Marcelo Cezar Pinto[1]

1 UNILA

Abstract

The development of biological databases has become indispensable for scientists in the field of bioinformatics. Biological data encompass a wide variety of complex information as well as large datasets. Scientists have used many tools to deal with those data (e.g. spreadsheets). However, these tools are not suited to integrate data from different sources and do not make data easily readable. Therefore, the development of biological databases is essential to achieve readability and integration of different analysis tools. Thus, this work in progress aims to develop a complete website for biological collections (databases) that allows the management of curators, researchers, assistants and guests with various projects running in parallel, integrated with a Geographic Information System. At this moment, the Fish Collection of UNILA and the management of research staff and projects are almost finished. For this website application, the backend side is coded using a Python-based framework called Django (available at <https://www.python.org/> and <https://www.djangoproject.com/>) integrated with PostgreSQL database. The frontend side is made with Bootstrap toolkit as well as AJAX with JavaScript Object Notation (JSON) to retrieve data dynamically. Because the data access is based on the role of the user (curator, researcher, assistant or guest), the query process was made by presenting the data filtered in HTML tables using the DataTables plug-in (available at <https://datatables.net/>). An important functionality of the website is the process of data importation and exportation. This tool is composed of modules, where each one will deal with some external pattern, like SpecieLink (available at <http://splink.cria.org.br/>). For example, the exporting tool will have a module to import data from a Microsoft Access database called "Coleção de Peixes da UNILA" and another one to export to FishBase website (available at <http://www.fishbase.org/>). At this moment, all the tables regarding the UNILA Fish Collection have been modeled with the curatorship of an expert.

Funding: This project is registered as PID202-2015, PID495-2016, PID575-2016, and PID1038-2017 and is partially funded by PIBITI-UNILA.

EXPLORATION OF REPRESENTATION OF POLYPEPTIDE CHAINS IN VECTORIAL MODELS FOR GENOMIC AND PROTEOMIC ANALYSIS

Ricardo Voyceik[1], José Miguel Ortega[2], Camilla Reginatto de Pierri[3], Letícia Graziela Costa Santos[3], Roberto Tadeu Raittz[3]

1 UFMG
2 UFMG, LABORATÓRIO DE BIODADOS
3 UFPR

Abstract

The volume of information in genes and proteins databases continues to increase exponentially, so vast amounts of biological information are available on public databases. Whilst some extent and requires increasing computational capacities to analysis performance, the ability to analyze this information has not been developed in the same way. Even with the today data mining techniques, useful for the treatment of large amounts of information, are limited in the exploration of biological data based on sequences, because they are unstructured and redundant data. Therefore, in this context, alignment free analysis of sequences was shown as a promising technique for analysis of proteomes and genes. With the aim of producing a mathematical model of representation of knowledge bases of genes or proteins, which can be manipulated by mathematical properties, in what it is possible to calculate the similarity between the amino acid sequences quickly, allowing the clustering by the biological characteristics given by the amino acids contained in the sequences, we propose an approach for gene and proteomic representation in a structured and reversible vector model, which has the potential to improve the capacity of analysis and data mining of biological data derived from protein sequences. Farther, we make available a graphic model to visualize this form of representation. The proposed model consists on the decomposition of sequences in sliding-windows analysis, in order to generate vectors that represent each amino acid at the sequences. As results, in the analyzes of the tests performed with the comparative among the techniques available in sequence analysis, in the aspects of clustering and similarity measures, the proposed method showed to have equivalent sensitivity, with the advantage of providing a substantially superior computational performance. We also show that the linearization of the matrices is a vector model that allows algebraic operations between the represented units, giving feasibility to operations such as geometric averages, cosine distance, centroid definitions, principal component analysis and the reduction by projection of the complete vector model for orthonormal basis. The reduction by projections to orthonormal basis down to 211 attributes, maintained the sensitivity of the model above 99%. As the vector model is storage in floating-point numbers, the data can be manipulated by GPUs with fast and precise parallel processing and possibility of clustering as well as machine learning techniques can be approached directly in the proposed model, without having to extract his characteristics or conversions, which opens new horizons in the data mining in Bioinformatics using math analysis tools.

Funding: CAPES

StatGraph: a statistical tool to analyze biological networks

Suzana de Siqueira Santos[1], Daniel Yasumasa Takahashi[2], Andre Fujita[1]

1 IME - USP
2 DEPARTMENT OF PSYCHOLOGY AND NEUROSCIENCE INSTITUTE, PRINCETON UNIVERSITY, PRINCETON, UNITED STATES OF AMERICA

Abstract

Several biological systems can be modeled by graphs (networks), which represent the interactions/relationships among the elements of the system. Examples of systems represented by graphs include gene regulatory networks, protein-protein interaction networks, and brain functional networks. Understanding how components interact with each other and the changes that occur in the system under certain conditions is a major concern in several fields of biology. One challenge of biological system studies is dealing with their variability. For example, the coactivation between regions of the brain changes across time, and gene regulatory networks are not identical even among individuals that share the same phenotype. Therefore statistical methods that model the randomness/variability of the graph structure are essential to analyze biological networks. However the available computational tools to analyze graphs lack statistical methods (frequently they do not include any statistical test or include tests for only few specific features of the graph). Here we present an R package, namely statGraph, that includes several statistical methods to analyze graphs, such as a test to compare the structure among two or more sets of graphs, a test to verify whether two sequences of graphs are correlated, a model selection approach and a procedure for parameter estimation. All methods included in statGraph are based on the spectrum of the graph (set of eigenvalues of the adjacency matrix), which is associated with several properties of the graph, such as number of walks, cliques and diameter. The methods were validated through simulation experiments, which show that they behave as expected for graphs with sufficient number of vertices.

Funding: S.S.S. was partially supported by CAPES and FAPESP (2015/21162-4). A.F. was partially supported by FAPESP (2013/03447-6, 2015/01587-0, 2016/13422-9), CNPq (304876/2016-0), NAP eScience-PRP-USP.

What is the pig's order? Dealing with the ragged hierarchy of NCBI Taxonomy

Testsu Sakamoto[1], Lab Biodados[2]

1 UFMG. LABORATÓRIO DE BIODADOS.
2 UFMG

Abstract

Any biological data are tightly linked to taxonomy and several bioinformatics tools use this information to answer biological questions regarding the species classification, diversity and evolution. One important taxonomic source is the NCBI Taxonomy. In this database, all sequence accessions deposited on INSDC are associated with their respective species which they belong; and the catalogued species are organized in a hierarchical structure that attempts to illustrate their evolutionary relationship. The hierarchical structure of NCBI Taxonomy is useful in retrieving the taxonomic lineage and classification since it uses taxonomic nomenclature and rank (class, order, family etc.) to name and classify the nodes comprising it. Though several bioinformatics analyses work with this taxonomic source, many of them face some difficulties because its hierarchy is of ragged type. As result, taxonomic lineages in the hierarchy are comprised of different number of nodes and some taxonomic ranks could be missing in some lineages. For instance, the taxonomic lineage of Sus scrofa (pig; taxid: 9823) does not have a node of subphylum, superclass, subclass or order ranks. Furthermore, the existence of nodes without a rank assigned (referred as "no rank") also contributes on generating ragged hierarchy. To address this issue, we developed an algorithm that takes the tree structure from NCBI Taxonomy and generates a balanced taxonomic tree. To create a balanced hierarchy, the algorithm firstly attempts to assign a taxonomic rank to a "no rank" nodes. Then, the algorithm creates/deletes nodes throughout the tree making it balanced. The algorithm also creates a name for the new nodes by borrowing the names from its ranked child or, if there is no child, from its ranked parent node. The new hierarchical structure was named Taxallnomy and it contains 29 hierarchical levels correspondent to the 29 taxonomic ranks used in the NCBI Taxonomy database. From Taxallnomy, user can obtain the complete taxonomic lineage with 29 nodes of all taxa available in the NCBI Taxonomy database. Taxallnomy is applicable to several bioinformatics analyses that depend on the data from NCBI Taxonomy. In this work, we demonstrated its applicability by embedding taxonomic information of a specified rank to a phylogenetic tree; and by making metagenomic profile according to a rank. The algorithm was written in PERL and all resource for Taxallnomy database can be accessed at biodados.icb.ufmg.br/taxallnomy.

Funding: FAPEMIG, CAPES

All purpose word pairing tool: Easy interaction networks for clinical data.

Thaynã Nhaara Oliveira Damasceno[1], Euzébio Guimaraes Barbosa[1]

1 UFRN

Abstract

Considering that the hospital environment daily generates a large volume of data, and as well as the volume of published biomedical research, and therefore the underlying biomedical knowledge base, it's necessary to use specialized tools are able to transform data into information that influence in a positive way to decision-making in relation to clinical practice. The increasing of data available in the databases of organizational data, as well as the need of techniques that are most appropriate for its analysis has facilitated the emergence of new techniques for Data Mining, aiming at a better analysis of these. The DM can be defined as the process of discovery of patterns and relationships considered relevant within large data sets. As an extension of the Data Mining area, Text mining as being an application of computer systems involving both hardware and software in the textual analysis of documents. An algorithm called Integrate Paired Tool (IPT) was developed using the languages JavaScript, HTML, CSS, R, Perl and Shell Scripting. This tool enable quick tools to create interactive Gephi input files to plot Interaction Network from data described is lists. The algorithm is designed for a large variety of data, but it has a large impact to simplify data retrieved from clinical databases. The IPT uses techniques of Data Mining and Text Mining for analysis of Clinical Data, and these data can be aggregated by any professional in the multiprofessional team in health, not restricted to only one subarea. The tool has performed the analysis taking into consideration their own data supplied by the user. After pairing, the tool generates two files that can be displayed in the tool Gephi, one with the nodes and another with the edges of the network. Gephi is an open-source software for network visualization and analysis. Gephi allows the end user to operate, analysis, use of filters, the manipulation of data, as well as cluster and export of data from any type of networks. We hope that our tool could be further extended and used to analyze data from Pubmed queries in the future due to its powerful way to extract meaningful data from complex data files.

Funding: UFRN/CAPES/PROPESQ

MCSM-PPI v2: predicting the effects of mutations in protein-protein binding affinity from sequence and structural features

Willy Garabini Cornelissen[1], David B. Ascher[2], Douglas E.v. Pires[1]

1 INSTITUTO RENÉ RACHOU, FUNDAÇÃO OSWALDO CRUZ
2 DEPARTMENT OF BIOCHEMISTRY AND MOLECULAR BIOLOGY, UNIVERSITY OF MELBOURNE, MELBOURNE

Abstract

Recent studies have been showing that a large proportion of disease mutations, including inherited genetic diseases as well as cancer-related mutations, occur at protein-protein interfaces (PPIs). In fact, edgatic mutations (those affecting interactions) seem to be overrepresented in diseases. Considering their importance in biology and for public health, the ability of predicting the effects of missense mutations on PPIs from protein sequence and structural evidences has become an important step towards understanding their role in diseases as well as an important tool for protein engineering. Different computational approaches have been proposed in recent years for this purpose, although much less attention has been devoted to mutations in protein-protein interactions and how they interfere with the binding affinity of the complex.

Here we present mCSM-PPI v2, an optimised machine learning-based method that relies on sequence and structural information to quantitatively predict the impact of single-point missense mutations in the affinity of protein-protein complexes. The improved method encodes distance patterns between atoms in a feature vector aiming to capture both the geometry and physicochemical properties of protein residue environments as well as distance patterns between residues on a sequence level. These signatures are then used as evidence for algorithms to learn from mutations with available experimental thermodynamic data.

mCSM-PPI v2 has been retrained, tested and optimised on low redundancy data sets derived from the SKEMPI database. Our method obtained a Pearson's correlation coefficient of up to 0.80, considerably more accurate than alternative approaches (BeAtMuSiC = 0.47; Li et al. = 0.58; Moal et al. = 0.73;), showing that using sequence and structural information in combination was advantageous. mCSM-PPI v2 was also able to identify hot-spots on PPIs via computational alanine scanning as well as optimising peptide affinity via in silico saturation mutagenesis.

We believe mCSM-PPI v2 will be scalable, robust, quantitative approach for analysing large data sets of mutations in PPIs that may aid in optimizing protein-protein binding and modulation as well as understanding deleterious mutations and their relation with diseases.

Funding: Instituto René Rachou, Fundação Oswaldo Cruz, Belo Horizonte

8 | Systems Biology and Networks

Evaluation of WGCNA and NERI methods for prioritization of pathways associated to schizophrenia spectrum disorders

Arthur Sant'anna Feltrin[1], Ana Carolina Tahira[2], Sérgio Nery Simões[3], Helena Brentani[2], David Correa Martins Jr[1]

1 UFABC
2 USP
3 INSTITUTO FEDERAL DO ESPÍRITO SANTO

Abstract

Using two gene expression data related to schizophrenia, we proposed a new approach consisting of combining the results of two network analysis algorithms: Weighted Gene Correlation Network Analysis (WGCNA) and Network-Medicine Relative Importance (NERI). Considering the differences between the two methods, our hypothesis is that both are capable of producing compelling results related to different aspects of schizophrenia's biological pathways; therefore, are complementary to each other. For that, we used replication and enrichment analysis using public databases. WGCNA uses gene expression from two groups to build co-expression pairwise correlation matrices, using connectivity parameters for evaluation of the network. NERI also uses expression data, but its network construction is based on the integration of PPI databases, gene expression, and a previously chosen seed genes list; the network analysis are based on shortest ranking path and relative importance calculation. We conducted an enrichment analysis using DAVID for the identification of partial biological function of each result, as well a replication and MSET analysis (for GWAS, transcriptome, methylation and de novo mutation databases related to schizophrenia) to appraise the replication and accuracy of our new approach when compared with each method in separate. The WGCNA module represents a final network of 435 and 300 genes on BAHN and KATO expression data. The enrichment analysis of this group using ppi modules leads to 88 genes across 10 hyper-represented human modules (adj.p<0.05), mostly involving immunological process. By using NERI, the final gene list was 150 genes for both BAHN and KATO with the enrichment analysis leading to modules related to glutamate receptor signaling, apoptotic process, neurotrophin and MAPK pathways. Both methods achieved statistical relevant replication results (p<0.05), but with one gene shared between both methods results. In the MSET analysis, NERI was capable for achieve meaningful results for the methylation and de novo mutation databases; whether our proposal of combining both results achieved better results for these two databases and for transcriptome, also increasing the number of candidate genes of each list. Our study suggests that using both methods in combination could be a promising approach for establishing a group of modules and pathways related to schizophrenia (or any complex disease).

Funding: FAPESP Ref.: 2014/10488-3; Universidade Federal do ABC

Niji: Analysis on the origin of biological systems using KEGG Pathways

Carlos Alberto Xavier Gonçalves[1], José Miguel Ortega[2]

1 UFMG
2 UFMG. LABORATÓRIO DE BIODADOS

Abstract

The Kyoto Encyclopedia of Genes and Genomes (Kegg) contains hundreds of pathways representing biological systems involved in metabolism, signaling, diseases and several other topics. These pathways are described in XML files and are graphically depicted in image files within Kegg's database. Kegg also contains data on clusters of orthologues, with which it is possible to obtain the taxonomic distribution of any given gene present in those pathways; by knowing all the organisms that contain a certain gene, it is possible do determine the lowest common ancestor (LCA) to those organisms, allowing us to infer the clade of origin of that gene. By using a local database containing LCA information for all genes on Kegg and also Kegg's automated programming interface (API), we generated colorized Kegg Pathways for the Homo sapiens in a way that each gene box's color is a representation of that gene's LCA; thus, genes that originate on the same clade are colorized with the same color. This allowed us to analyze how biological systems evolved over time. We also utilized Python scripts to recreate each pathway in graph objects, using the information contained in the XML files, and applied the LCA data to discover if the pathways were formed from a single connected component, or if they evolved from multiple subsystems that eventually coalesced. Of the 314 Kegg maps analyzed, 35 did not contain any edge information on Homo sapiens. We encountered 46 systems that reach full connectivity on the Homo sapiens, meaning no elements on those systems are disconnected at the most recent clade. Of these, 15 (32%) evolved on a single growing component, with new elements connecting directly to previously existing entities, whereas 31 (68%) evolved from multiple coalescing subsystems. Interestingly, six (13%) of the 46 fully-connected pathways are entirely ancient, with all elements dating back to the origin of eukaryotes, while there are seven (15%) maps containing up to early animals genes (from Metazoa through Bilateria). The remaining 33 (72%) maps have genes originated within the chordates. Most of these pathways reached full completeness within the Euteleostomi (modern fishes) clade, and some are as recent as the placental mammals (Theria and Eutheria clades). We created an online platform for consultation of these data, called Niji (the Japanese word for rainbow), that is available at: biodados.icb.ufmg.br/niji

Funding: CAPES, CNPq, FAPEMIG

Use of data mining for Onco-targets to analyze Breast Cancer through the construction of Ontology Networks

Edgar Lacerda de Aguiar[1], Lissur Azevedo Orsine[2], José Miguel Ortega[3]

1 CEFET-MG
2 UFMG
3 UFMG, LABORATÓRIO DE BIODADOS

Abstract

Studies indicate that by the end of 2017 more than 23 million patients will develop some type of cancer. Among the various types of cancer, breast cancer has the most impact among women and one of the highest mortality rate. Breast cancer has high biological heterogeneity, which implies a high diversity of molecular forms which are associated with distinct subtypes and distinct drug targets. This high range of variations in the biological entities involved in disease pathology impacts directly on diagnosis and treatment. Because of these facts this work aims to mine the possible genes related to breast cancer, from different databases (DBs), Cancer Genome Atlas (TCGA), COSMIC, KEGG and to relate the genes to database Gene Ontology (GO) with its molecular functions, biological processes, cellular components, thus inferring the main ontologies associated with breast cancer. Initially there was an interpolation of genes between BDs CGA and COSMIC after cured genes were mined and crossed with the top mutated genes of breast cancer. The genes were cured with the BDs UniProt, NCBI and KEGG, focusing on DB KEGG Pathway, which was used to search for the genes of the various types of cancer. With the UniProt valors Id of cured Genes there was a search for Ontologies in the GO database. The initial results weren't very satisfactory due to high specificity and high granularity of ontological terms. Better treatment of the data and a new methodological approach to the Ontological terms was necessary. The ontological terms of GoSlim, which are terms and healed with a median specificity were used. Several enrichment analyzes were performed comparing breast cancer genes with genes responsible for breast development. Through these analyzes it was possible to note that approximately 10% of the breast cancer genes were found in the group of genes responsible for breast development. In the analysis of biological processes of the gens were associated 42% in metabolic process, 39% in response to stimulus and 33% in developmental process. In the Molecular function 33% in binding, catalytic activity 34% and 10% receptor activity. The analyzes show a significant correlation between many biological processes and molecular functions encountered. This work is important for a better understanding of breast cancer and the genes responsible for breast development, through better analysis of biological processes, molecular functions using data mining and ontology network, to aid in the search for Onco-targets.

Funding: Cefet-Mg, Ufmg

Understanding Immunosenescence through a Systems Biology Approach

Fernando Marcon Passos[1], Helder Takashi Imoto Nakaya[2]

1 USP

Abstract

The remodeling of the immune system that comes with age, known as immunosenescence, contributes to an increased susceptibility in elderly to infectious diseases, cancer, autoimmunity and decreased vaccines response. This remodeling is a complex and multifactorial process and, until now, there is little understanding of the molecular mechanisms involved. Several studies tried to understand and identify which genes and signaling pathways are involved in the ageing of our immune system. However, none has yet done a comprehensive analysis of a large amount of transcriptomic data of healthy subjects in a wide age spectrum. In this project we aim to create a predictive model for the biological age of the immune system using machine learning methods. For that we will perform a meta-analyses of microarray transcriptomic data available in the GEO public repository. We selected 29 studies containing 435 blood samples that had subject's age information. First, we will identify genes that are differently expressed between age groups, through the statistical method LIMMA. With such genes we can discover the gene signatures that are related with immunosenescence throughout a pathway enrichment analysis. In this step, we will also perform a coexpression analysis to build gene networks related to immunosenescence. This will be done using the CEMiTool, a tool developed in our laboratory that allow us to identify gene modules and sub-modules associated with a particular phenotype. The next step is to create a predictive model of the biological age of the immune system. We will use dimensionality reduction and feature selection algorithms, like PCA and the FSelector package, to select genes that optimize the predictive power of the model. Then, we will use various algorithms of machine learning, such as Support Vector Machine and Neural Networks, to create age group classification and age regression models. These models will then be validated with blood samples from children and elderly, which will be provided by the Liverpool School of Tropical Medicine. Once the model has been validated, genes used in machine learning algorithm as well as the regulation profile and co-expression of gene networks discovered in the meta-analysis will be used to understand the mechanisms of activation and deactivation of the genes they are related to immunosenescence.

Funding: FipFarma

Identification of brain regions associated with neurodevelopment

Grover Enrique Castro Guzman[1], Maciel Calebe Vidal[1], João Ricardo Sato[2], André Fujita[3]

1 USP
2 UFABC
3 INSTITUTE OF MATHEMATICS AND STATISTICS - USP

Abstract

Initial studies using resting-state functional magnetic resonance imaging on the trajectories of the functional brain network from childhood to adulthood found evidence of functional integration and segregation over time. The comprehension of how healthy individuals' functional integration and segregation occur is crucial to enhance our understanding of possible deviations that may lead to brain disorders. Recent approaches have focused on the framework wherein the functional brain network is organized into spatially distributed modules that have been associated with specific cognitive functions. Here, we tested the hypothesis that the clustering structure of brain networks evolves during development. To address this hypothesis, we defined a measure of how well a brain region is clustered (network fitness index - NFI), and developed a method to evaluate its association with age. Then, we applied this method to a functional magnetic resonance imaging data set composed of 397 males under 31 years of age collected as part of the Autism Brain Imaging Data Exchange (ABIDE) Consortium. As results, we identified two brain regions for which the clustering change over time, namely, the left putamen and the right frontal pole. Since the NFI is associated with both integration and segregation, our findings suggest that the two identified brain regions play a role in the development of brain systems.

Funding: FAPESP (2015/01587-0, 2016/13422-9), CNPq (grants 304020/2013-3), CAPES, and NAP-eScience-PRP-USP

Investigation of the replication-transcription conflicts in Trypanosoma brucei through computational dynamical models

Gustavo Cayres[1], Marcelo S. da Silva[1], Marcelo S. Reis[1], Maria C. Elias[2]

1 INSTITUTO BUTANTAN
2 LECC-CETICS, BUTANTAN INSTITUTE

Abstract

In the context of Molecular Cell Biology, DNA replication consists on the process of duplicating the genetic material of a cell. This process can start multiple times during the S-phase of cell cycle, at specific genomic regions named "replication origins". However, the triggering frequency of each origin and the dynamics of its respective replisomes are subject to variations along S-phase. Moreover, the influence of the collisions between these replisomes and the DNA polymerase (DNAP) on the overall duration of the S-phase is unknown. Therefore, our objective in this work is the development of computational dynamic models to test whether replisome/DNAP collisions have relevant impact on the S-phase dynamics in various protozoa species in the kinetoplastida group. We started this investigation with Trypanosoma brucei, the pathogen behind the sleeping sickness. The proposed model consists in a Markov chain whose transition function can be estimated using heterogeneous data (e.g., the distribution of replication origins and the transcription sites of each chromosome) obtained from the literature and also from wet-lab experiments carried out at our lab. This information was organized into a relational database and a model simulator was implemented in Python. Unknown parameters such as the transcription initiation frequency and number of available replication origin sites were evaluated through a comprehensive Monte Carlo sampling on a search space constrained by the biological feasibility of the values obtained in a given simulation. Initial results with T. brucei strain 927 showed that a causal response to replisome/DNAP collisions (e.g., through the ATM/ATR signaling pathways) is not necessary to accomplish DNA replication within the S-phase required time that is reported in the literature. Currently, we are applying this methodology into other protozoa such as T. cruzi, the parasite that causes Chagas disease. Therefore, we expect to elucidate how differences on the replication dynamics of these organisms accounts for differences in the genomic architecture that are observed in kinetoplastids.

Funding: CNPq and grants #2013/07467-1, #2016/17775-3, and #2016/50050-2, São Paulo Research Foundation (FAPESP).

A global feature selection algorithm for the model selection step in the identification of cell signaling networks

Gustavo Estrela de Matos[1], Lulu Wu[2], Vincent Noel[3], Marco Dimas Gubitoso[4], Carlos Eduardo Ferreira[4], Junior Barrera[4], Hugo A. Armelin[3], Marcelo S. Reis[3]

1 CENTER OF TOXINS, IMMUNE-RESPONSE AND CELL SIGNALING, INSTITUTO BUTANTAN, INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
2 CENTER OF TOXINS, IMMUNE-RESPONSE AND CELL SIGNALING, INSTITUTO BUTANTAN
3 INSTITUTO BUTANTAN
4 INSTITUTO DE MATEMÁTICA E ESTATÍSTICA, USP

Abstract

In the context of cell signaling network identification, model selection is the choice of a dynamic model from a set of possibilities; the chosen model should be the most suitable one according to a given cost function (e.g., curve-fitting optimization). If these possibilities are defined by differences in the chemical species and/or reactions that compose each of them, then a feature selection procedure could be carried out to accomplish the model selection. Recently, it was proposed a method to carry out model selection through examination of interactome databases. However, such databases typically yield huge search spaces during the feature selection procedure; hence, only a greedy sequential approach could be explored so far. Therefore, there is a need for development of efficient global feature selection methods to tackle this hard combinatorial optimization problem. In this work, we introduce a new global feature selection method, which may be used to assist the model selection step during the identification of cell signaling networks. This method, called Parallelized U-Curve Search (PUCS), relies on the fact that the chain costs of the Boolean lattice induced by the search space are decomposable in U-shaped curves; this latter phenomenon is due the curse of dimensionality, that is, the impact the lack of samples brings to the cost function as the number of considered features increases. To implement and evaluate the PUCS algorithm, we used featsel, a framework for benchmarking of feature selection algorithms and cost functions. To compute the cost function (i.e., the fitness of a candidate model), we are employing the Signaling Network Simulator (SigNetSim), a tool for building, fitting, and analyzing mathematical models of molecular signaling networks. Initial results with synthetic data showed that PUCS outperforms golden standard algorithms in feature selection such as the Sequential Forward Floating Search (SFFS). Currently, we are applying PUCS into the model selection of real-data signaling networks extracted from the Kyoto Encyclopedia of Genes and Genomes (KEGG) interactome database. We expect that this new feature selection method will become a critical asset to identify fully predictive dynamic models, which in turn will help researchers to unveil intricate molecular mechanisms underlying cell phenotype changes due extracellular stimuli.

Funding: CAPES, CNPq and grants #2013/07467-1 and #2016/25959-7, São Paulo Research Foundation (FAPESP)

Extraction of features using topological measures of complex networks

Isaque Katahira[1], Eric Augusto Ito[1], Fábio Fernandes da Rocha Vicente[1], Fabricio Martins Lopes[1]

1 FEDERAL TECHNOLOGICAL UNIVERSITY OF PARANÁ

Abstract

The feature extraction methods are important for the study of the large amount of data produced by the high-performance sequencing techniques. The dimensionality reduction methods have been used to summarize the most significant characteristics of a data source. The goal is to represent a great volume of data from its characteristics, minimizing the information loss. Thus, the current research proposes a model of feature extraction based on the theory of complex networks for the representation of biological sequences. The proposed model consists on sequence mapping in graphs, in which the vertices are the segments of a sequence and the edges are defined by their structural organization (neighborhood). These edges are weighted by the pair occurrence frequency of adjacent segments in the input sequence. Then, topological measures of graphs are extracted: motifs, degree, minimum degree, maximum degree, standard deviation, cluster coefficient, average path length, proximity and intermediation. These measures compose a feature vector that represent a sequence, which is used to classify the input sequences. Coding and non-coding transcripts of nine species were used in order to verify the suitability of the proposed method, using the algorithms Random Forest, Naive Bayes, LibSVM and J48. A 10 fold cross-validation was performed to evaluate the predictors. The maximum accuracy for the coding transcripts identification was reached by the LibSVM with 100%; followed by J48 with 99.43%; Random Forest with 99.38%; and Naive Bayes with 93.64%. The best index related to the accuracy for the identification of non-coding transcripts was reached by Naive Bayes with 93.12%; followed by Random Forest with 81.43%; J48 with 79.41% and LibSVM with only 6%. The predictor that obtained the best accuracy average between the classification of coding and non-coding was the Naive Bayes with 93.38%. The results indicate the validity of the proposal, considering that the extraction of topological characteristics of complex networks got significant values ??of accuracy, which can be extended to the classification of other biological sequences like DNA and amino acids.

Funding: CAPES, Fundação Araucária

An overview of etanol tolerance in Saccharomyces cerevisiae through systems biology and differential expression analysis

Ivan Rodrigo Wolf[1], Lauana Fogaça(department Of Bioprocess And Biotechnology. São Paulo State University[2], Leonardo Nazário de Moraes[3], Rafael P. Simões[4], Lucilene Delazari Dos Santos[5], Rejane M. T.grotto[6], Guilherme Targino Valente[4]

1 FCA-UNESP
2 UNESP
3 DEPARTMENT OF BIOPROCESS AND BIOTECHNOLOGY. SÃO PAULO STATE UNIVERSITY
4 UNESP - STATE USP
5 CENTER FOR THE STUDY OF VENOMS AND POISONOUS ANIMALS , SÃO PAULO STATE UNIVERSITY
6 DEPARTMENT OF BIOPROCESS AND BIOTECHNOLOGY, SÃO PAULO STATE UNIVERSITY, BOTUCATU

Abstract

The bioethanol production contributes to the sustainable development and the life's quality improvement. The main bioethanol production process is through the first generation technology, which Saccharomyces cerevisiae is the most widely used organism. Unfortunately, ethanol is toxic to S. cerevisiae in higher concentrations, limiting the bioethanol production. Tolerance to ethanol is a complex feature, and it is poorly understood, then the conventional methods have been unsuccessful in attempting to understand their mechanisms. Here we experimentally determined ethanol tolerance for five yeast strains (S288c, BY4741, BY4742, SEY6210, X2180-1A and BMA64-1A) and the unsupervised learning was used to classify the strains as high (HT) or low (LT) tolerant. RNA and proteins were further extracted for treatment (maximum ethanol exposure) and control (without ethanol exposure) conditions and submitted for sequencing and/or mass-charge quantification. The gene transcripts showing significant differences (FDR<0.05) between treatment and control were considered as differentially expressed (DE); a common set of 270 genes was found up regulated in treatment while 80 were down regulated. The extracted protein was submitted to mass spectrometry and a total of 18 protein-coding genes with fold-change>1, where considered up regulated (7 for HT, 7 for LT and 4 common to both HT and LT). Interestingly, 2 chaperones where coincidently found up regulated in transcript and protein data. The functions of differentially expressed genes have already been observed in previous studies. However, this is the first time it was observed synergistically in one experiment. A co-expression (CoEx-net) network was created based on transcripts abundance data for each strain. Differences between CoEx-net and our previous protein-protein interaction (PPI-net) are evident considering topological characteristics as the degree, density, diameter, and assortativity; it reflects different meanings between systems biology layers analyzed. The clustering analysis

How the Ebola infection happens and since when?

Elisson Nogueira Lopes[1], Lissur Azevedo Orsine[2], Iara Dantas de Souza[3], Tetsu Sakamoto[4], Rodrigo Juliani Siqueira Dalmolin[3], José Miguel Ortega[4]

1 UFMG
2 UFRN
3 UFMG. LABORATÓRIO DE BIODADOS.

Abstract

The Ebola virus (EBOV) is an enveloped, filamentous virus, and contains a negative-sense RNA genome. EBOV belongs to Filoviridae family and is the causative of a devastating disease, with a mortality rate of about 50-90%. The first symptoms developed by infected patient are fever, malaise and muscle pain, and could be followed by bleeding and organ failure. While Ebola initially targets macrophages and dendritic cells it is able to infect almost all cells types with exception of lymphocytes. The Ebola has been proposed to attach multiple plasma membranes and after that the viral glycoprotein induces uptake via macropinocytosis. The process is dependent on the action of cell surface proteins. After uptake into macropinosomes, particles travel to compartments where the viral glycoprotein is cleaved and fused to membranes, what results on the release of the viral compartments in host cytoplasm. We looked at the infection mechanism of EBOV and collected the host proteins known by now to participated direct or indirect on infection. This mining approach comprised 52 host proteins. With them, we built a pathway to represent the Ebola's cycle and interaction with host proteins. We also analyzed the homologous of each collected human protein along the taxonomic tree to infer their clade/epoch of origin. The results of the evolutionary origin analysis allowed to infer that the virus could infect even vertebrates, suggesting that animals such as fish and amphibians could be infected and retransmit the virus to other hosts such as man. Moreover we analyzed four GEO datasets for gene expression after Ebola infection, characterizing the enrichment of GO processes along the time-course of infection. Initially processes involving cellular checkpoint and DNA metabolism were enriched, followed by several other processes. In conclusion, Ebola infection happens with interaction with recent proteins in the membrane, interacts with more ancient proteins along its intracellular path and later on with more recent ones as the virus connects with proteins implicated in the immune response, and the pathway construction helps to add context to the time-course modulation of gene expression.

Funding: CAPES, FAPEMIG.

An integrated omics using Petri Net approach to the characterization of genetically modified yeast for second generation ethanol production

Lucas Miguel de Carvalho[1], Renan Pirolla[2], Gabriela Vaz de Meirelles[3], Leandro Vieira Dos Santos[2], Fabio Cesar Gozzo[4], Gonçalo Amarante Guimarães Pereira[5], Marcelo Falsarella Carazzolle[6]

1 UNICAMP
2 CNPEM - CTBE
3 LGE - UNICAMP
4 UNICAMP - IQ
5 BRAZILIAN BIOETHANOL SCIENCE AND TECHNOLOGY LABORATORY CTBE, BRAZILIAN CENTER FOR RESEARCH IN ENERGY AND MATERIALS CNPEM, BIOLOGY INSTITUTE - UNICAMP
6 BIOLOGY INSTITUTE - UNICAMP, NATIONAL CENTER FOR HIGH PERFORMANCE COMPUTING CENAPAD-SP/UNICAMP

Abstract

Brazil is one of the biggest producers of ethanol in the world, a pioneer in the ethanol industry. However, the country is already facing a major limitation imposed by the first-generation ethanol production technology, in which the sugarcane juice is converted by ethanol using industrial yeast Saccharomyces cerevisiae. Therefore, a new alternative approach has been proposed, called second generation, which is based on lignocellulosic residues of sugarcane (bagasse and straw) for ethanol production using recent methodologies for biomass deconstruction that generates soluble sugars, majority represented by glucose and xylose. One of the biggest challenges of this technology is the development of genetically modified industrial yeast that can not only produce ethanol from glucose as usual, but also from xylose that represents 15% to 45% of the lignocellulosic material. Several works have developed xylose-fermenting yeast using different exogenous genes and genetic engineering approaches, but always resulting in very low yield and productivity mainly caused by unbalanced redox potential and metabolic bottleneck. Nowadays two metabolic pathways for the consumption of pentoses are known: oxido-reductase (OXR) pathway, identified in fungi, and xylose isomerase (XI) pathway frequently found in bacteria. Using genetic engineering tools is possible to insert these two metabolic pathways into the industrial yeast in order to enable it to consume pentoses with different fermentative performances. The combination of omics data (transcriptomic, proteomic and metabolomic) and bioinformatics analysis is an essential step for a better understanding of this system. Moreover, biological models based on experimental datasets can recreate several biological aspects in different conditions using a combination of integrated omics and computational simulations. Quantitative stochastic models of molecular interaction networks can be expressed as Stochastic Petri Nets (SPNs), a mathematical formalism developed in computer science. In this work we

Cancer immunology of Cutaneous Melanoma: A Systems Biology Approach.

Mindy Muñoz[1], Thiago Dominguez Crespo Hirata[2], Pedro de Sá Tavares Russo[2], Melissa Lever[2], Helder Takashi Imoto Nakaya[3]

1 COMPUTATIONAL SYSTEMS BIOLOGY LABORATORY - FACULDADE DE CIÊNCIAS FARMACÊUTICAS,USP
2 USP

Abstract

Cutaneous melanoma is a melanocyte skin cancer and it is one of the most aggressive tumors in humans. It causes a great number of deaths worldwide, and in Brazil approximately 1,300 melanoma patients die each year. The Cancer Genome Atlas (TCGA) database contains genomics, epigenomics and transcriptomics data from 471 samples of skin cutaneous melanoma (SKCM). A few studies have applied systems biology approaches to investigate melanoma progression. However, they failed to integrate several layers of "omics" data in order to elucidate the mechanisms by which melanoma cells become resistant to the immune system. We propose here to perform an integrative omics analysis with the SKCM data available in TCGA. For this, we will utilize established models coupled with hub detection algorithms. The identification of hub genes can help us to unravel the role of immune system in SKCM progression.

Funding: CAPES

Group-Directed Biasing Effects on Topological Properties of PPI Networks

Paulo Burke[1], Luciano da Fontoura Costa[1]

1 IFSC - USP

Abstract

Complex networks have increasingly been used for representing and analyzing biological systems such as protein-protein interaction, metabolism, and gene regulation. However, most these networks are substantially incompletely sampled as a consequence of experimental difficulties. So, it becomes important to investigate to which extent such incompleteness can bias the network representations, especially regarding the estimation of several topological properties. Though some related studies have been reported in the literature, they mostly focus on uniform sampling biases, therefore not including situations in which one or more groups of nodes or edges are, by their biological nature, differently affected by sampling. This case is henceforth called group-directed biasing. Indeed, this situation is commonly found in biology, such as in the case of proteins with high content of exposed apolar amino acids bias effect on yeast two-hybrid (Y2H) assays. The present work aims at investigating such situations, by using simulations. More specifically, we build diverse model networks which are biologically more plausible (e.g. Barabási-Albert) in Protein-Protein Interaction (PPI) networks, select subgroups of nodes and/or edges which may or may not share topological characteristics, and derive respective sampled versions of these networks with sampling biasing specific to groups of nodes. Then, several topological measurements are obtained for these networks and compared to the original models. In this way, we provide insights about the effect of different types of group-directed biasing on the accuracy of the estimation of topological features of complex networks.

Funding: FAPESP, CNPq, CAPES

CEMiTool: Coexpression Modules Identification Tool

Pedro de Sá Tavares Russo[1], Gustavo Rodrigues Ferreira[1], Lucas Cardozo[1], Matheus Carvalho Bürger[1], Raúl Arias-carrasco[2], Sandra Regina Maruyama[1], Thiago Dominguez Crespo Hirata[1], Diógenes Saulo Lima[1], Fernando Marcon Passos[1], Kiyoshi Ferreira Fukutani[1], Melissa Lever[1], João Santana Silva[1], Vinicius Maracaja Coutinho[2], Helder Takashi Imoto Nakaya[3]

1 USP
2 UNIVERSIDAD MAYOR

Abstract

The analysis of co-expression gene modules can help uncover the mechanisms underlying diseases and infection. We present a fast and easy-to- use Bioconductor package named CEMiTool that unifies the discovery and the analysis of co-expression modules. Using the same real datasets, we demonstrate that CEMiTool outperforms existing tools, and provides unique results in a user-friendly html report with high quality graphs. Among its features, our tool evaluates whether modules contain genes that are over-represented by specific pathways or that are altered in a specific sample group, as well as integrate transcriptomic data with interactome information, identifying potential hubs on each network. We successfully applied CEMiTool to over 1,000 transcriptome datasets, and to a new RNA-seq dataset of patients infected with Leishmania, revealing novel insights of the disease's physiopathology.

Funding: FAPESP

Integrative networks analysis based on RNAseq data to elucidate a presence of B chromosome

Rafael Takahiro Nakajima[1], Ivan Rodrigo Wolf[2], Guilherme Targino Valente[3], Rodrigo de Oliveira Almeida[2], Rafael P. Simões[3], Cesar Martins[1]

1 IBB-UNESP
2 FCA-UNESP
3 UNESP - STATE USP

Abstract

B chromosomes occur in about 2000 species, including animals, insects and plants. Several works have been conducted with the aim of understanding their distribution, frequency, transmission mechanisms, structure and origin. Cichlid fish receive great scientific interest, since many species are under rapid and extensive adaptive radiation. Astatotilapia latifasciata is one of the species of African cichlids that presents B chromosomes. In this species, Bs, although heterochromatic, present genes with high integrity and interfere in the transcriptional profile of the cells. Thus, the present work aims to characterize possible candidate genes of A. latifasciata specific tissues to elucidate the influence of the presence of B chromosomes in specific metabolic pathways from data obtained from RNASeq. For this purpose, networks were constructed by concatenating co-expression and protein-protein interaction networks, which obey the expected degree-distribution for biological networks. Ontologically enriched domains were extracted from the network for important biological processes to be compared with differential expression data of mRNAs in gonads. Results of the intersection between networks and differential expression, presented an important role in the regulation of cellular activity, mainly to an anti-inflammatory response, the presence of cellular membrane processes and components that may be related to the defense mechanism.

Funding: FAPESP

Using machine learning to cluster genes and tissues according to their functions through gene co-expression

Thaís de Almeida Ratis Ramos[1], José Miguel Ortega[2], Vinicius Maracaja Coutinho[3], Thaís Gaudencio[4]

1 UFRN
2 UFMG, LABORATÓRIO DE BIODADOS.
3 UNIVERSIDAD MAYOR
4 UFPB

Abstract

The creation of gene expression encyclopedias possibilities the understanding of genes groups that are co-expressed in differents tissues and comprehend gene clusters according to their functions. The advent of machine learning, with unsupervised methods without needing to define the number of clusters a priori on the clustering process, is possible to map large data sets. This would be the first step to understanding the performance of transcription factors in the regulation processes of gene expression. The purpose of this work is to evaluate genes coexpression by tissue and function through gene expression data. For that, were tested 3 databases: Uhlen, Fantom and Encode. As pre-processing data, four normalization types were tested and adopted the combination of two of them: Transcripts Per Kilobase Million (TPM) and base-2 log. In the clustering process we use 2 machine learning algorithms: K-means and Hierarchical implemented in an online tool calling CORAZON (Correlation Analyses Zipper Online). To select the best number of clusters were used: Bayesian information criterion (BIC) followed by the derivative of the discrete function and Silhouette. Furthermore, in the Hierarchical we test eight linkage criterions and adopted the Ward's method. The first database with 32 tissues had an optimal number of clusters equal to 9, the second with 56 tissues, 11 clusters and the last with 13 tissues, 7 clusters. We observed that hierarchical method and K-means generated exactly the same clusters, only a few had some slight variation among their components. However, we can observe groups related to glands, cardiac tissues, muscular tissues, tissues related to the reproductive system and in all three groups are observed with a single tissue, such as testis, brain and bone-narrow. The same uniformity behavior was found in the functional analysis of the gene groups after clustering. In the first database were analyzed 44594 genes in 9 clusters; 21080 genes in the second database grouped into 11 clusters and the third database grouped 42355 genes into 10 clusters. In relation to the genes clusters, we obtained several clusters that have specificities in their functions: detection of stimulus involved in sensory perception, reproduction, synaptic signaling, nervous system, immunological system, system development, and metabolics. These results are preliminary but show the methods potential and the possibilities of analyzing transcription factors in the regulation process of these genes, making possible the evolutionary history study of genes and the biological system regulatory map.

Funding: UFRN

Dynamical model of the Ras-mediated AP-1 activation in mouse Y1 adrenocortical tumor cells.

Vincent Noel[1], Marcelo S. Reis[1], Matheus H.s. Dias[1], Cecilia S. Fonseca[1], Francisca N.l. Vitorino[1], Layra L. Albuquerque[1], Fabio Nakano[2], Julia P.c. da Cunha[1], Junior Barrera[3], Hugo A. Armelin[1]

1 INSTITUTO BUTANTAN
2 ESCOLA DE ARTES, CIÊNCIAS E HUMANIDADES, USP, BRASIL
3 INSTITUTO DE MATEMÁTICA E ESTATÍSTICA, USP

Abstract

The K-Ras-driven mouse adrenocortical tumor cell line Y1 displays a surprising association of phenotypic traits, i.e., high basal levels of activated K-Ras in starved cells and induction of cell cycle arrest upon stimulation by FGF2. In addition, ectopic expression of the dominant negative mutant Ras-N17 reduced activated K-Ras basal levels and eliminated cell cycle arrest by FGF2. We are working to uncovered the molecular basis of this unexpected phenomenon by building a dynamical model of the Ras-MAPK signaling pathway in Y1, and it's subsequent activation of the AP-1 complex. This model has been designed using SigNetSim, a web platform for modeling signaling networks being developed by our team. We notably used its functionality of building hierarchical models to produce a modular network, and simplify the reuse of already existing models. We started our K-Ras molecular switch model by reusing a published Ras model. We then added another GEF to reproduce the high basal K-Ras activation observed in Y1, and a Ras dominant negative which expression would reduce the K-Ras basal level. With these modifications, our model was able to reproduce experimental observations from our team. Then, we linked our K-Ras model to a published model of MAPK pathway, by making its activation dependent on Ras activation. With this additional module, we were able to reproduce experimental observations of MAPK activation. We finally added the translocation of MAPK to the nucleus, and its expression of the AP-1 complex. We were able to start building a model to describe the unusual behavior of Y1 cells. Our model reproduces the behavior of K-Ras, MAPK, and AP-1 activation in starved cells, serum stimulated cells, and Serum+FGF2 stimulated cells. We are presently studying the cell cycle activation by AP-1, and the additional stress produced by FGF2 stimulation, to incorporate it in our model and be able to reproduce the cell cycle blockage. We are also planning to improve the conditions covered by our model to reproduce the observations on FGF-resistant Y1 sublines.

Funding: This work was supported by grants #12/20186-9, #13/07467-1, and #13/24212-7 of the São Paulo Research Foundation (FAPESP).

SigNetSim : A web platform for building and analyzing mathematical models of molecular signaling networks

Vincent Noel[1], Marcelo S. Reis[1], Matheus H.s. Dias[1], Lulu Wu[2], Amanda S. Guimares[3], Daniel F. Reverbel[3], Junior Barrera[3], Hugo A. Armelin[1]

1 INSTITUTO BUTANTAN
2 CENTER OF TOXINS, IMMUNE-RESPONSE AND CELL SIGNALING, INSTITUTO BUTANTAN
3 INSTITUTO DE MATEMÁTICA E ESTATÍSTICA, USP

Abstract

Molecular biology is experiencing a revolution, in one part thanks to new technologies to measure and perturb biological systems in vitro, and also due to the growing importance of mathematical modeling which enables us to understand biological mechanisms in a more profound way. However, a crucial point in this transforming field is the need to provide completely new tools, which should be computationally efficient, versatile, and compatible. To this end, we developed SigNetSim, a web platform to create, simulate, adjust and analyze biochemical reaction models. As a web platform, it does not require powerful devices and is usable on multiple systems. It is designed to be installed on computation servers, with most of the work executed server-side. Users can create and edit biological models by describing the species and the reactions in the model. Reactions can be defined by their kinetic law and associated parameters, or by their mathematical formula. To assist the creation of large models, users can also include submodels as part of they models. This also encourage and simplify the reuse of existing models. Models can also be annotated, using the MIRIAM guidelines. SigNetSim can perform model simulation for time-series and steady states. Users can also look for dynamical properties such as bifurcations in the steady states of the systems, using continuation techniques. The platform includes a simple database to store experimental data, which can be used to simulate models according to a set of initial conditions and compare the results with experimental observations, or to fit models to reproduce observations, using a parallelized simulated annealing algorithm. This algorithm allows users to estimate missing parameters, even in large systems. SigNetSim is using community standards to store most the work done by the users. Models are stored in SBML models, and can be imported/exported to Biomodels database from the interface. Simulations are stored using SEDML, and can be easily exported to online repositories such as JWS Online. Data from the database can be exported using NUML format. Whole project can be saved in one file using the COMBINE archive standard. The compatibility with these standards ensure the reproducibility of the research work, and help collaborating even using different tools. Finally, SigNetSim is distributed under AGPL3 license, and its core library under GPL3 license. It is available at signetsim.org and on GitHub.

Funding: This work was supported by grants #12/20186-9, #13/07467-1, and #13/24212-7 of the São Paulo Research Foundation (FAPESP) and fellowships from CNPq.

BioNetStat: A differential network analysis tool to biological data

Vinicius Jardim Carvalho[1], Suzana de Siqueira Santos[2], Andre Fujita[2], Marcos Silveira Buckeridge[1]

1 USP
2 IME - USP

Abstract

The networks theory is an important way to model and understand the interactions diversity of biological systems, considering from cells organelles to the whole biosphere. The dynamic of systems structure, such as the changes in the interactions among the system elements, is an inherent trait of those systems. To represent each one of the many states assumed by a system we can use networks. In this sense, there is a wide range of tools proposed to compare those networks. However, none of them are able to compare structural characteristics among more than two networks. Considering that systems generally assume more than two states, we developed a statistical tool to compare more than two networks and highlight key players in the process studied. The main proposition of this study was to compare correlation networks using traits that are based on graph spectra (eigenvalues set of adjacency matrix), such as the spectral distribution. This measure is associated with other traits of networks, such as the number of walks, diameter, and cliques, and it is a better characterization of graphs than other classical measures of networks theory. In addition to spectral distribution, we also compare networks by spectral entropy, degree distribution, and nodes centralities. To verify the performance of tool we used a tumoral cells genes expressions data set. In the case studies, we used two data sets, the same gene expression data and a plant metabolites concentrations data. The method proposed, called BioNetStat, was implemented in R package with a user interface for people that do not programming. We verified that the method is efficient to distinguish more than two networks. However, the increase in networks number and the decrease in sample unities reduce the statistical power of methods. The method proposed brings a potential time economy, doing a single analysis to compare more than two networks rather than compare them by pairs. The method highlighted sets of variables with a central role in biological systems that were not highlighted in other studies where only gene expression or metabolic concentration were analyzed. In that way, we proposed another way to find traits (variables) that distinguish cancer types genes expression or organ plants metabolisms. Furthermore, the highlighted variables allow us to create hypotheses about its role in the process studied, bringing new finds about the systems operation mechanisms.

Funding: FAPESP, CAPES

Both mechanism and age of duplications contribute to biased gene retention patterns in plants

Hugo Rody Vianna Silva[1], Luiz Orlando de Oliveira[2]

1 UNIFESP
2 UFV

Abstract

Funding: CAPES

Computational gene expression environment by agent-based mRNA translation modeling

Anton Semenchenko[1], Guilherme Oliveira[2], A. P. F. Atman[3]

1 CENTRO UNIVERSITÁRIO NEWTON PAIVA
2 VALE TECHNOLOGY INSTITUTE
3 CENTRO FEDERAL DE EDUCAÇÃO TECNOLÓGICA DE MINAS GERAIS

Abstract

Funding: CENTRO FEDERAL DE EDUCAÇÃO TECNOLÓGICA DE MINAS GERAIS

BLASTing NGS data with CrocoBLAST

Ravi Jose Tristao Ramos[1]

1 CEITEC - CENTRAL EUROPEAN INSTITUTE OF TECHNOLOGY

Abstract

Funding: CEITEC-Central European Institute of Technology, Masaryk University,

CeRNAs in plants: computational approaches and associated challenges for target mimic research

Alexandre R. Paschoal[1], Irma Lozada-chávez[2], Douglas Silva Domingues[1], Peter F. Stadler[3]

1 UTFPR - PPGBIOINFO
2 UNIVERSITY OF LEIPZIG, GERMANY
3 UNIVERSITY OF LEIPZIG. GERMANY

Abstract

Funding: Self-funding and DAAD (GER)

GeNICE: A Novel Framework for Gene Network Inference by Clustering, Exhaustive Search, and Multivariate Analysis

Ricardo de Souza Jacomini[1], David Correa Martins Jr[2], Felipe Leno da Silva[3], Anna Helena Reali Costa[3]

1 USP
2 UFABC
3 ESCOLA POLITÉCNICA DA USP

Abstract

Funding: FAPESP, CNPq, CAPES

SnoRNA and piRNA expression levels modified by tobacco use in women with lung adenocarcinoma

Natasha Jorge[1], Gabriel Wajnberg[2], Carlos Gil Ferreira[3], Benílton Carvalho[4], Fabio Passetti[1]

1 FIOCRUZ - IOC
2 FIOCRUZ-IOC
3 D'OR INSTITUTE FOR RESERACH AND EDUCATION, RIO DE JANEIRO
4 DEPARTMENT OF STATISTICS, STATE UNICAMP

Abstract

Funding: CAPES, FAPERJ, CNPq, FAPESP

PacBio assembly of a Plasmodium knowlesi genome sequence with Hi-C correction and manual annotation of the SICAvar gene family

Juliana Assis[1], Mary Galinski[2], Jéssica Kissinger[3]

1 UFMG
2 EMORY VACCINE CENTER, YERKES NATIONAL PRIMATE RESEARCH CENTER - EMORY UNIVERSITY
3 CENTER FOR TROPICAL AND EMERGING GLOBAL DISEASES - UNIVERSITY OF GEORGIA

Abstract

Funding: Federal funds from the National Institute of Allergy and Infectious Diseases; National Institutes of Health, Department of Health and Human Services (Contract No.HHSN272201200031C) and the National Center for Research Resources (ORIP/OD P51OD011132). This study was also financially supported by the National Institutes of Health (R01 AI06775-01) to KGLR, the University of California, Riverside (NIFA-Hatch-225935) to KGLR and Institute Leadership Funds from La Jolla Institute for Allergy and Immun