Proceedings X-Meeting 2015

Editor: AB^3C

December 4, 2015

Conference Program

Submitted Paper Abstracts 235

Index of Authors

1 | Organizing Committee

AB3C President: Glória R Franco (UFMG)

AB3C Vice President: Alan M Durham (USP)

BSB Chair: Sérgio Campos (UFMG)

AB3C Secretaries :

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

AB3C Financial Department :

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

Poster Session Organizers :

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

Paper Submission Organizers :

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

Local Committee :

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

2 | Introduction

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

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

3 | Abstracts

ProClaT, a new bioinformatics tool for in silico protein reclassification: case study of DraB, a protein coded from the draTGB operon in Azospirillum brasilense

Elisa Terumi Rubel, Roberto Tadeu Raittz, Nilson Antonio da Rocha Coimbra, Michelly Alves Coutinho Gehlen, Fabio de Oliveira Pedrosa

Federal University of Paraná - Curitiba

Abstract

Azopirillum brasilense is a plant-growth promoting nitrogen-fixing bacteria that is used as bio-fertilizer in agriculture. Since nitrogen fixation has a high-energy demand, the reduction of N2 to NH4+ by nitrogenase occurs only under limiting conditions of NH4+ and O2. Moreover, the synthesis and activity of nitrogenase is highly regulated to prevent energy waste. In A. brasilense nitrogenase activity is regulated by the products of draG and draT. The product of the draB gene, located downstream in the draTGB operon, may be involved in the regulation of nitrogenase activity by an, as yet, unknown mechanism. A deep in silico analysis of the product of draB was undertaken aiming at suggesting its possible function and involvement with DraT and DraG in the regulation of nitrogenase activity in A. brasilense. In this work, we present a new artificial intelligence strategy for protein classification, named ProClaT. The features used by the pattern recognition model were derived from the primary structure of the DraB homologous proteins, calculated by a ProClaT internal algorithm. ProClaT was applied to this case study and the results revealed that the A. brasilense draB gene codes for a protein highly similar to the nitrogenase associated NifO protein of Azotobacter vinelandii. This tool allowed the reclassification of DraB/NifO homologous proteins, hypothetical, conserved hypothetical and those annotated as putative arsenate reductase, ArsC, as NifO-like. An analysis of co-occurrence of draB, draT, draG and of other nif genes was performed, suggesting the involvement of draB (nifO) in nitrogen fixation, however, without the definition of a specific function.

Meeting the Global Thirst for Bioinformatics Training

Global Organisation of Bioinformatics Learning, Education and Training

GOBLET

Abstract

Year on year, the demand for bioinformatics training and skills has steadily increased. The breadth of audiences soliciting bioinformatics training has also increased, expanding from primarily wet-lab scientists to include high-school teachers and students, as well as junior faculty and seasoned academics. These trends are being experienced globally. The Global Organisation of Bioinformatics Learning, Education and Training (GOBLET) is a network of bioinformatics trainers and initiatives. GOBLET's core mission is to provide a global, sustainable support and networking structure for capacity development of bioinformatics trainers and trainees. This includes a training portal, allowing trainers to share materials, tools and techniques, guidelines and best-practices documents, and resources. In addition, GOBLET is i) fostering the international community of bioinformatics/computational biology trainers through networking events, ii) facilitating bioinformatics capacity development across the globe, particularly through its train-the-trainer and train-the-teacher initiatives, and iii) developing best-practices standards and guidelines for bioinformatics training. Over the past year, GOBLET has put forward numerous bioinformatics training programs to meet the growing global demand for such training. Presented here are some of the GOBLET training enterprises, which have taken place at various ISMB conferences, science teacher association conferences and other venues around the world. We share lessons learned in organizing and presenting these sessions, and discuss the impact these sessions have had on further outreach efforts.

Standardizing sequence analysis methods on the web: working on the Bioinformatics Platform at the Fiocruz-BA

Alisson Fonseca, Artur Lopo, Luciano Silva

FIOCRUZ-BA

Abstract

The Bioinformatics Platform-BA [RPT04D] was created in July 2012 on the Technology Platforms Network FIOCRUZ (FIOCRUZ platforms) and provides services in bioinformatics to assist in the research developed at the CPqGM-FIOCRUZ / BA. In order to publicize the services and standardize the methods of analysis, we have developed a page in the virtual learning environment of FIOCRUZ-BA (AVA FIOCRUZ, `http://ava.bahia.fiocruz.br/`) based on the Moodle platform (Moodle.Org). Access to the RPT04D page is available to visitors through: AVA FIOCRUZ → Training → Resources and procedures of Bioinformatics Platform - BA [RPT04D] and offers the following contents: resources, procedures, articles in bioinformatics, sequencing technologies, sources for training in software development and programming languages, database and useful links. Among resources we list the hardware and software of RPT04D and other units of FIOCRUZ-BA, which are available for shared use, and ways of access: local and / or remote (text terminal, graphic terminal or web - http, ftp). In addition to a network server, some paid software is available: DNAStar Lasergene Core Suite v. 11 CLC Main Workbench v. Vector NTI Express 6.9 and 1.1.2. Among the procedures that have been standardized are: search sequences (BLAST), sequences assembly (DNAStar), primer design (DNAStar), multiple alignment (ClustalX), typing and analysis of resistance mutations of HBV and HIV (HIVSeq & HBV Seq , Stanford University), and restriction fragment cloning (DNASTAR) software. Open source software (free software) or software based on the web were evaluated to allow us to make available additional resources unavailable in paid software. The creation of this page does not limit the work of researchers and students with advanced knowledge, but provides an opportunity for those less experienced to solve simple problems using bioinformatics tools autonomously.

Comparative genomics of three species of the genus Echinococcus

Lucas Luciano Maldonado, Juliana Assis, Flávio Marcos Gomes-Aráujo, Izinara Rosse, Natalia Macchiaroli, Marcela Cucher, Mara Rosenzvit, Guilherme Oliveira, Laura Kamenetzky

IMPAM-UBA-CONICET, CEBio, Instituto Tecnológico Vale, Belém-BR

Abstract

Echinococcus canadensis is a platyhelminth parasite member of the class Cestoda which keeps close phylogenetic relationship with Echinococcus granulosus and Echinococcus multilocularis which are involved in hydatid infections of humans and animals. In South America three species of Echinococcus spp. have been reported E. granulosus sensu stricto (G1, G2 and G3 genotypes), E. canadensis (G6 and G7 genotypes) and E. ortleppi (G5 genotype), all belonging to the complex Echinococcus granulosus sensu lato. High quality genomic DNA was extracted and two paired-end libraries were constructed and sequenced by Illumina technology. Reads were trimmed using trimmomatic and de novo assembled using SPAdes. Metrics of the assembly were evaluated by QUAST and the completeness of the genome was validated by CEGMA. Gene Models were generated by MAKER and accuracy statistics for the gene set predicted were evaluated with EVAL. Proteins were annotated using InterPROScan and by generating orthology groups using OrthoMCL. Orthology groups were constructed by using proteomes from representative species of the phylum Platyhelminthes and by representative eukaryotic proteomes. In order to identify SNPs we mapped reads on E. multilocularis and E. granulosus (G1) using Bowtie2 and variants were detected using samtools and vcftools. After filtering out low quality SNPs, variants were annotated using snpEFF.Metrics of the assembly showed a high quality genome (i. e size 115Mb, N50 74.555Kb, # contigs 9332, 11499 genes). As expected; high synteny was found between the genomes of the three species of Echinococcus. Specific cestode orthology groups have been detected. Regarding SNPs analysis, the rate of synonymous substitution is higher than the nonsynonymous rate but we found more missense mutations between E. canadensis (G7) and E. granulosus (G1) than between E. canadensis (G7) and E. multilocularis. This result was unexpected since E. canadensis (G7) and E. granulosus (G1) belong to E. granulosus sensu lato. Orthology analysis from E. canadensis (G7) high quality genome has demonstrated high synteny between related species. Specific cestode proteins were selected for further analysis. E. canadensis presents more SNPs in comparison with E. granulosus than with E. multilocularis. The effect of genetic variants are more likely to behave as synonymous mutations, without generating significant changes into amino acid sequences. Further analysis have to be performed in order to confirm these results. The knowledge of this new genome provide information for comparative genomics allowing to adapt diagnosis tools to each epidemiological situation and helping to understand the biological differences observed between species.

BIOMARKERS STUDY OF NEPHROTOXICITY CAUSED BY GENTAMICIN THROUGH GENE EXPRESSION

Marina Grossi Stellamaris Soares, Sarah Silva, Mariana Nunes, Leonardo Almeida, Anete Valente, Carlos Tagliati

Federal University of Minas Gerais, Belo Horizonte, Brazil., Federal University of Alfenas, Alfenas, Brazil, Federal University of Juiz de Fora, Governador Valadares, Brazil

Abstract

Drug-induced nephrotoxicity is one of the most frequently observed effects in the early preclinical phase of drug development. The effects of nephrotoxicity are commonly discovered later due to lack of sensitivity of in vivo methods to evaluate this effect. Therefore, researchers have tried to develop in vitro alternative methods for the early identification of toxicity. Identification of drug-induced gene changes is critical to providing insights into molecular mechanisms and to detecting renal damage. Gentamicin, for example, is a widely used aminoglycoside antibiotic that causes nephrotoxicity. In the present study, LLC-PK1 cells were exposed for 24 h to gentamicin concentrations of 4 (low), 8 (medium), and 12 (high) mM, according to MTT tests, to evaluate gene expression. A literature survey was conduced to identify genes associated with the development of nephrotoxicity. A panel of genes was selected based on gene expression changes in multiple published studies. Due to the limited base of study for the cell model in this work, the search for sequences of mRNA encoding proteins that had been previously associated with kidney damage was researched in the databases of the National Center for Biotechnology Information - NCBI (USA). The primers were obtained using the Primer BLAST (NCBI) program, based on the sequences of selected transcripts. RNA was extracted from the cells, and RT-PCR was performed to evaluate expression profiles of the selected genes. Among the analyzed genes, four genes proved to be highly up-regulated in cells exposed to the nephrotoxin: HAVcr1 (hepatitis A virus, cell receptor 1), CASP3 (caspase1), ICAM1 (intracellular adhesion molecule 1), and EXOC3 (exocyst complex component 3). According to the obtained results, it can be suggested that these genes can be used as early in vitro biomarkers for the identification of nephrotoxicity. The establishment of genomic markers is a promising tool for evaluating nephrotoxicity and will be useful in the development of safer drugs.

Identification of small deletions within human exons in Asian and European genomes using trancriptome data.

Gabriel Wajnberg, Nicole de Miranda Scherer, Carlos Gil Ferreira, Fabio Passetti

FIOCRUZ/Oswaldo Cruz Institute, INCA

Abstract

Insertions and deletions (INDEL) are examples of alterations in the DNA sequence. If one INDEL is located within the coding region, it can produce transcripts with modifications in splice sites, the encoded amino acids or frame shift. The 1000 genomes project can be used as source of data to search for INDELs in different populations. We used an innovative method to search for INDELs: usage of transcriptome data to identify small deletions up to 99 nucleotides in length within human coding exons. Here, we present preliminary data from the analysis of 12 genomes from the 1000 genomes: 6 East Asian Ancestry (EAS) and 6 European Ancestry (EUR). A total of 291,049 small deletions were identified and 96,046 may cause frameshift. We detected deletions previously identified only by the 1000 genomes project (57), only annotated in the dbSNP (67) and by both approaches (5). For example, we detected previously annotated deletions in the dbSNP in the following human genes: DHFR (rs144629981) in EUR, TIFA (rs5861095) in EAS and DNAI2 (rs140867882) in both EUR and EAS. We were able to detect novel unannotated small deletions: 2,004 in more than 3 EAS genomes (exclusive to EAS), 1,451 in more than 3 EUR genomes (exclusive to EUR) and 17,435 in more than 50% of all genomes analyzed. Some frameshift small deletions are predicted to affect important protein domains in the following cancer-associated genes: ICAM1 (exclusive to EAS), CASP8 (exclusive to EAS) and CDK2 (EAS and EUR). In conclusion, we present preliminary data in which we used transcriptome data to identify deletions previously described in the dbSNP and the 1000 genomes project, and to detect novel unannotated deletions in human genes. Financial support: FIOCRUZ, INCA/MS, CAPES, Fundação do Câncer, FAPERJ and CNPq.

Characterization of pathogenicity islands of 15 genome of Corynebacterium pseudotuberculosis biovar equi

Yan Patrick de Moraes Pantoja, Adonney Allan Oliveira Veras, Pablo Henrique Caracciolo Gomes de Sá, Vasco Azevedo, Adriana Ribeiro, Artur Silva, Rommel Ramos

Federal University of Pará, Federal University of Minas Gerais

Abstract

Corynebacterium pseudotuberculosis is a Gram-positive pathogen and the main cause of Caseous Lymphadenitis (CL), infectious disease of worldwide occurrence found mostly in goats and sheep that at present has no cure and is responsible for great economic losses due to the reduction in the production of meat, milk and wool and can also more rarely infect humans. This bacterium has large genomic regions called genomic islands, which are acquired horizontally from other species that carry genes encoding one or more virulence factors. According to the content, the genomic islands may be known as pathogenicity islands, that hosting clusters of virulence genes that mediate the adhesion, colonization, invasion, immune system evasion, and toxigenic properties of the acceptor organism. Currently, there are several techniques and computational tools to perform the prediction of pathogenicity islands in prokaryotic genomes. Therefore, was used the software Genomic Island Prediction Software - Gipsy to make the prediction of the islands of the 15 genomes of C. pseudotuberculosis. For visualization of predicted pathogenicity islands and other relevant information about the target genome is used so-called genome browsers. Among the available, it was decided to use the JBrowse, which is web-based and provides the viewing and the genomic comparison without the need for software installation or configuration. It uses an AJAX interface, fully dynamic that helps preserve the user's sense of location by avoiding discontinuous transitions, instead offering smoothly animated panning, zooming, navigation, and track selection. Furthermore, a software was developed to automating the process of inserting files in JBrowse order to facilitate and expedite this process. The identification of pathogenicity islands in the 15 genomes of C. pseudotuberculosis along with their characterization is of paramount importance to facilitate the search for this group of genes that encode to virulence factors of the bacteria, thus contributing to a better characterization of the organism and subsequent control of their pathogenicity.

Analysis of HPV regulatory elements by k-mer index approach

Taina Raiol, Lucas Akayama, Luciana Montera

Oswaldo Cruz Foundation, Federal University of Mato Grosso do Sul, Federal University of Mato Grosso do Sul

Abstract

The human papillomaviruses (HPVs) are the primary etiologic agents of cervical cancer and have been associated to the development of other anogenital and non-melanoma skin cancers. They are strictly epitheliotropics and are further classified in either mucosal or cutaneous depending on viral tropism, which is determined by the tissue type where each HPV genotype has been more frequently or exclusively detected. It has been suggested that the enhancer located within the viral regulatory region (LCR) is directly involved in the cellular tropism, however there are very few evidences about the molecular determinants of this complex event. In order to investigate potential regulatory regions involved in the viral tropism, we developed a k-mer index method for comparison of LCRs from these two HPV groups. This approach was implemented in python to search for all k-mers from a set of LCRs of each specific group. To address the sequence composition among different genotypes and groups, it is allowed at most m mismatches during the search process. The search is implemented as a force brute algorithm, however, once a search of an k-mer from a sequence lcr_i is done, identical k-mers from other input sequences do not need to be searched. Besides, the processes are accelerated by using hash and set data structures from python. The tool is available at http://pintado.facom.ufms.br/hpv from witch one can perform a search considering at most two datasets. Not only all k-mers from each set are presented, but also a Venn Diagram representing the intersection among the datasets, if its exists. The LCR sequences were extracted from all HPV genome sequences (26 cutaneous and 13 mucosal) publicly available at Genome database (NCBI). The comparison was performed using k-mers ranging from 15 to 30 and 0 to 7 mismatches. Interestingly, using k-mer 18 and allowing up to 7 mismatches for cutaneous HPVs and 6 for mucosal HPVs, we could detect 159 different sequence patterns exclusively for cutaneos group and 327 for mucosal group. In addition, we could detect patterns that match to known binding sites present within the LCR, such as C/EBP and E2F1. Our preliminary results indicate that there are exclusive sequence patterns for each group of HPV, which may be signatures for specific sets of transcription factor binding sites involved in HPV epithelial specificity.

Maximum entropy: an effective strategy to discriminate anomalous regions in bacterial genomes

Gesiele Barros-Carvalho, Marie-Anne Van Sluys, Fabrício Lopes

University of São Paulo, Federal University of Technology- Paraná

Abstract

The number of complete bacterial genome sequences have increased recently. As a result, there is a need for new alternatives to quickly identify and extract information about the organism. Knowing the genomic plasticity of bacteria and that genome diversification during evolutionary processes provide specific and important features that distinguishes closely related lineages, we aimed to develop a simple and effective methodology based on maximum entropy (ME) to quickly reveal anomalous regions. Thus, directing the analysis to regions that should be prioritized during the study. This methodology needs just to genome in fasta format as an input, a priori, and consists of two main steps: sliding window and maximum entropy calculation. The main output are a graph with ME distribution in the genome, their coordinates and the anomalous regions predicted. Besides, to annotated genomes, the proposed approach may provide a protein list present in each selected region and warning whether some selected regions contain ribosomal. The methodology has been applied to two bacteria (Xanthomonas axonopodis pv. citri 306 and Xanthomonas campestris pv. campestris ATCC 33913), which have the anomalous regions well known. The results were compared with others available methodologies as Alien Hunter, HGT-DB, Islander, IslandPath and SIGI-HMM. The ME approach showed higher efficiency and F-score values than competing methods used for the same objective. Moreover, maximum entropy may be considered a fast computational option for a holistic analysis of the genome, because it spent less than seven minutes, on computer with Intel Core i3-540 Processor and 8 GB RAM, to analyze each Xanthomonas genome. Therefore, the proposed methodology based on maximum entropy proved to be a good alternative of analysis for individual genomes, and enable to discriminate genomic regions in a simple and fast way, without relying on previous annotation and comparison with other genomes. Such strategy can provide a direction to researcher, in order to prioritize genomic regions to be explored in more detail and streamline the analysis time. Besides, it is the first time that the maximum entropy is used to analyze genome sequences. The ME can also be a new option in the early stages of analysis of newly sequenced genomes.

The complete mitochondrial genome of a brazilian carnivorous plant Utricularia reniformis (Lentibulariaceae): Insights into the evolution of an organellar genome of a specialized plant.

Yani C A Diaz, Saura R Silva, Cristine G Menezes, Vitor F O Miranda, Todd P Michael, Alessandro M. Varani

Univ. Estadual Paulista, Ibis Bioscience, Univ. Estadual Paulista, Câmpus Jaboticabal

Abstract

Carnivorous plants from the genus Utricularia (Lentibulariaceae) are distributed world-wide, comprised of approximately 235 species occurring across every continent except Antarctica. They are highly specialized plants with specialized leaves (traps) to capture prey and have both aquatic and terrestrial forms. In addition, the genus has some of the smallest nuclear genomes across the angiosperms, ranging from 79 to 561 Mbps (1C). Currently, little is known about the contents and organization of the Utricularia mitochondrial genome. Here, we report the first complete mitchondrial genome of the genus. Utricularia reniformis A. St.-Hil. (Lentibulariaceae), which is a terrestrial form endemic to the Atlantic Forest and restricted to the mountaintops of the southeastern coast of Brazil. The U. reniformis mitochondrial genome was sequenced on the Illumina MiSeq platform with a total of 10 million 2x300bp paired-end reads for a estimated mtDNA genome coverage of 500x. The genome was assembled with SPAdes v3.5.0 and Platanus v1.2.1 software, and manually annotated with the Mitofy and DOGMA pipelines. Based on comparison with the NCBI organelle database, the U. reniformis mtDNA is the third largest plant mitochondrial genome at 775,178bp. Its GC content is 43.99%, and it encodes 58 coding regions and 22 tRNAs. Using REPuter software, several repeats were identified that may be involved with different intra-molecular recombination events. Indeed, several smaller and circular molecules derived from the master mtDNA were detected using different assembly approaches and parameters, consistent with recombination. Fine-scale sequence analysis indicated that at least two recombinant sequence repeats of 24,910bp, and 23 bp in length divided the master mtDNA genome in four sub-genomic circles (SGC) with different sizes (SGC A: 653,976bp; SGC B: 121,203bp; SGC C: 603,385bp; and SGC D: 175,509bp). Another interesting feature of U. reniformis mtDNA is the acquisition of several regions from the chloroplast genome (cp). A total of two cytochrome related genes (petD and petB) and 53 pseudogenes were cp-derived. Finally, five genes encoding to a RNA-dependent RNA polymerase (RdRP) from the mitoviruses family were identified. Mitoviruses are the simplest viruses in that they are unencapsidated, and encode only a RdRP protein. Together, these results indicate that active recombination events and frequent gene transfer from the cp genome and mitovirus have shaped the mtDNA genome structure and organization.

OBO-RO Editor: Supporting development and integration of ontologies in the biomedical domain

Ricardo Cacheta Waldemarin, Cléver Ricardo Guareis de Farias

Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto - USP

Abstract

Currently, several domains of science are facing a huge increase in the quantity and variety of data types being produced and stored. In the biomedical domain, the advent and accessibility of new techniques has resulted in the production of a growing volume of heterogeneous data. In this scenario, conceptual modeling artifacts, such as ontologies, have been used to organize and integrate data from different sources in a coherent manner. The success of ontologies in the biomedical domain has led to the proliferation of such ontologies. However, the lack of standardization and alignment efforts between the developed ontologies represents itself an obstacle to the integration and use of these ontologies. The Open Biological and Biomedical Foundry (or, simply, OBO Foundry) was created as a collaborative experiment for the development and management of ontologies in the biomedical domain. The OBO Foundry aims at obtaining more correct, modular and integrable ontologies. In this sense, the OBO Foundry developed the Relation Ontology in order to provide formal definitions for a set of general purpose relationships widely used in the biomedical domain. A UML profile has been proposed to formally define in UML the different types of concepts and relationships provided by the OBO Relation Ontology, and to allow the creation of ontologies as UML models based on the set of definitions. The graphic notation provided by the profile facilitates the modeling and visualization of biomedical ontologies. Biomedical ontologies are often large and complex artifacts, which represents a limitation for their graphical representation and visualization. However, this limitation can be overcome by proper support of a modeling tool. Since the concepts are formally defined in the profile, such a tool can be developed to support automated reasoning in order to prevent syntactic inconsistencies in the ontologies being modeled. In this sense, this project aims at investigating the development of a modeling tool to support the construction of ontologies using the proposed UML profile. This modeling tool should provide support for automatic verification of models to ensure the syntactic consistency. Additionally, this project also aims at investigating the integration of ontologies developed using UML and ontologies developed using the OBO Foundry's ontology representation language, the OBO File Format.

THE DRAFT GENOME OF Fonsecaea multimorphosa CBS 980.96

Aniele C Ribas Leao, Vinicius Almir Weiss, Emanuel Maltempi de Souza, Vania A Vicente ACR Leao1, VA Weiss1, RR Gomes3, VA Vicente3, RT Raittz1, MZ Tadra-Sfeir, E Balsanelli, V Baura, MBR Steffens, EM Souza

Laboratory of Bioinformatics, Professional and Technological Education Sector - Federal University of Parana – UFPR, Department of Biochemistry and Molecular Biology – Federal University of Parana – UFPR, Department of Microbiology, Parasitology and Pathology - Federal University of Parana – UFPR

Abstract

Fonsecaea multimorphosa species belongs to Chaetothyriales order as a part of black yeasts clade, foundand in the environment and animal hosts. Pathogenic Chaetothyriales are known as causative agents of infections, such as subcutaneous mycosis, which leads to formation of muriformes cells, warts and high melanin production in the host. The high concentration of melanin serves as protection for the fungus, favoring resistance to therapeutic treatment. Records of the disease are located in most cases in the tropical and subtropical regions, commonly affecting rural workers, by direct contact with plants and soil contaminated with the fungus. The difficult diagnosis and treatment results in systemic infection, necrosis in the region with ringworm and amputation of the affected limb by the yeast. F. multimorphosa CBS 980.96, in contrast with other environmental strains from Brazil, was isolated from a cat brain abscess in Australia. Brain infections caused by filamentous fungi, such as F. multimorphosa are rare in animals, with leads the importance of having his genome elucidated and studied. The genome sequencing of F. multimorphosa CBS 980.96 was performed in the Illumina MiSeq platform generating ∼7 million of paired-end reads. The SPAdes assembler was used for the genome assembly and the FGAP software as a gap closure tool. The draft assembly was obtained resulting 288 contigs, 52% GC content, and size of 34MB. The dotplot alignment of the sequence under study was done by Mummerplot software, using the genome of the F. multimorphosa CBS 102226 as reference. The sequence of the mitochondrial DNA was also assembled with size of 27Kb, based on Fonsecaea pedrosoi and Exophiala dermatitidis mtDNA. The genome annotation was performed using the orfinder GeneMarK witch identified ∼47.000 genes and the soft wares BLAT and InterProScan for sequence similarity searches. Supported by: CAPES; CNPq; INCT – FBN.

COMPARISON OF BINNING SOFTWARES IN METAGENOMIC SEQUENCES FROM HUMAN GUT MICROBIAL

José Pilan, Agnes Takeda, José Rybarczyk-Filho

Institute of Biosciences of Botucatu – Univ. Estadual Paulista

Abstract

One of the first stages of Metagenomes analysis is binning, which consists on the analysis of nucleotide sequences (reads) obtained from the sequencing of a metagenome to classify the sequences in taxonomic groups. The correct taxonomic identification of sequences leads to a better characterization of a metagenome and helps in the genome assembly. There are several software tools that use different methods to attain these results. These methods are divided into two main categories: compositional tools and sequence alignment by similarity. These methods have differences in its processing time and the efficiency in reads identification. This work proposes the comparison between two methods of metagenome analysis using four samples taken from the project T̈he human gut microbial gene catalog established by metagenomic sequencing deep(ERR11305, ERR011323, ERR011277, ERR011259). The samples were submitted to the softwares Raiphy and Phymmbl, in which the first perform the alignment while the second uses both compositional and alignment methods. The results show that the main organims present in this sample are from families Eubacteriaceae, Halobacteroidaceae, Ruminococcaceae, Bacteroidaceae, Lachnospiraceae, Porphyromonadaceae, Clostridiaceae, Peptostreptococcaceae. Both compositional methods from Raiphy and Phymmbl softwares could identify all sequences of samples while the alignment method of Phymmbl succeeded only between 21.40% (ERR011277), 30.03% (ERR011259), 22.12% (ERR011323) and 38.84% (ERR011305). Moreover, the processing time required for the analyses was faster in compositional methods: Phymmbl required 1-4 days while RaIphy took 2-5h. These results suggest the potential of compositional methods as a first step of evaluation of the reads obtained from metagenome sequencing.

SEQUENCING AND ANALYSIS OF Derxia lacustris HL12

Sheyla Trefflich, Vinícius Almir Weiss, Arnaldo Glogauer, Dieval Guizelini, Roberto Tadeu Raittz, Shih-Yi Sheu, Wei-Cheng Huang, Wen-Ming Chen, Michelle Zibetti Tadra-Sfeir, Helisson Faoro, Valter Baura, Emanuel Maltempi Souza

Federal University of Paraná, National Kaohsiung Marine University, Laboratory of Microbiology, Department of Seafood Science

Abstract

Rhizobacteria are root-colonizing organisms that forms symbiotic relationships with many plants. These bacteria are of great importance in agriculture, since the shortage on nitrogen absorption directly impacts the plants's growth rate of, reducing agricultural productivity and leading to environmental problems with the use of harmful fertilizers. The biological nitrogen fixation is carried out by a wide range of bacteria. Derxia lacustris is a nitrogen-fixing bacterium witch belongs to the Betaproteobacteria class, Burkholderiales order, Alcaligenaceae family, and Derxia genus. This organism was isolated from water samples of a lake in Taiwan. Derxia lacustris HL-12 was the second bacterium classified in the Derxia genus , based on 16S ribosomal RNA gene sequence analysis, with the type strain specie Derxia gummosa . The genome of Derxia gummosa is deposited at GenBank and was assembled by 20 contigs. The genome of Derxia lacustris HL-12 was sequenced twice by Illumina MiSeq platform. The sequencing process resulted in a total of 3,532,510 paired end reads with an average length of 300 base pairs. The raw data were assembled by the software programs CLC, Velvet, Celera, Spades, and Masurca. The best assembly was achieved with Celera software, with 140 contigs and 69.99% of GC content. The MATLAB software was used for sort and genome calculations. Fgap software was used to close the gaps in the genome with the data from all the assemblies. Bowtie software was used for mapping the genome with raw data. Mummer software was used in the construction of dotplot between Derxia lacustris and Derxia gummosa. Fastqc software was used to analyze the quality of the raw data. Quast was the software used to compare the assembled data. Derxia lacustris had its genome fully closed, and showed one 16S-23S-5S ribosomal operon . The genome is under an ongoing annotation analysis for the identification of specific genes in metabolic pathways of the organism.

Challenges associated to the search of specific targets for drug development against Leishmania major

Larissa Catharina Costa, Carlyle Lima, Ana Carolina Ramos Guimarães, Nicolas Carels

Laboratório de Modelagem de Sistemas Biológicos, Centro de Desenvolvimento Tecnológico em Saúde (CDTS), Fundação Oswaldo Cruz (Fiocruz), Rio de Janeiro, Brasil.
Laboratoire de Biochimie Théorique, Institut de Biologie Physico-Chimique, Centre National de la Recherche Scientifique (CNRS), Université Paris 7, Paris, France.
Laboratório de Genômica Funcional e Bioinformática, Instituto Oswaldo Cruz (IOC), Fundação Oswaldo Cruz (Fiocruz), Rio de Janeiro, Brasil.

Abstract

Pharmaceutical industry has experienced a dramatic fall in productivity with an unprecedented requirement for investment in new drugs discovery and development. The number of new chemicals licensed by the FDA has been falling and the pharmaceutical industry has sought for new paradigms to reverse trend productivity to growth. The occurrence of false-positives in sequences annotation has been a recurring problem resulting in significant errors. In this report, we describe a method to cure false-positives based on the evaluation of the false-positives produced by AnEnPi since this error component is critical when one decides to invest in a putative specific enzyme for drug development. The methodology is characterized by a detailed functional analysis of enzyme specificity as well as analysis of secondary structure and catalytic domains, metabolic reconstruction and phylogenetic relationships in L. major to avoid false-positives cases. From 67 proteins sequences as specific, 15 parasite sequences produced homologous in the human genome. According to genomic hits coordinates, we recovered 80 proteins from their CCDS as annotated in Ensembl. After comparison between the coordinates of gene structure (exons + introns) from these CDSs and the coordinates of the protein sequence from human putative exons, we identified 14 putative cases of false-positives annotated with: (i) 8 enzymatic activities differing in the fourth digit; (ii) one case with an uncompleted EC number for the human enzyme; (iii) four cases that the sequence recovered from human genome were not annotated with enzymatic function and; (iv) one case that the EC number between the parasite and the host was different. Therefore, the procedure presented here follows a systematic organization that can be automated; it is suitable for host-parasite interactions involving a lower eukaryote as the parasite and a higher eukaryote as the host. Supported by: CNPq, PAPES/FIOCRUZ

Assembly, Annotation and Comparative Analysis of Mitochondrial Genome of two Asian Cattle breeds: Guzerá and Gir

Juliana Assis Geraldo, Izinara Rosse Maria Raquel Carvalho, Francislon Oliveira, Flávio Araújo, Marcos Vinícius Silva, Guilherme Oliveira

Centro de Pesquisas René Rachou, Universidade Federal de Minas Gerais, EMBRAPA, Instituto Tecnológico Vale

Abstract

This study aimed at assembling the mitochondrial genome (mtDNA) of two Asian cattle breeds (Bos indicus), Guzerá and Gir. Both of these breeds are the main milk production in Brazil. We assembled the mtDNA of both breeds, two animals for each breed (Gir 1-2 and Guzerá 1-2) to improve the understanding of the mtDNA molecular diversity within the Bos genus. The automatic annotation followed by high manual curation allowed the identification of the D-Loop no coding region, 13 protein-coding region, 22 tRNA, and two rRNA genes (12s and 16s) in all genomes analyzed. These results correspond to 37 typical mitochondrial genes in animals. Nucleotide composition, codon usage and identification of single nucleotide variations – SNVs were performed. When we looked for throughout the genome, we found highly variable regions and most substitutions at the 3rd codon position in coding sequences. Comparative analysis with other cattle genomes was performed and the high numbers of SNVs were found when compared to others breeds. These SNVs results highlighted the separation between the animals Gir and Guzerá 1 and Gir and Guzerá 2, these results suggesting that the Gir1 and Guzerá 1 carry the mtDNA from Zebu and the others carry mtDNA from Taurine. The phylogenetic reconstruction allowed to classify the genomes assembled in haplogroups and the point of origin of domestication of these animals was inferred. The complete mitochondrial sequences will be deposited in public databases to contribute to future work of bovine genetics. The results are important to highlight that changes in the mitochondrial genome (mtDNA) can lead to effects on cellular metabolism, and consequently on reproductive performance and/or milk or meat production and our find may contribute to the identification of new and more informative molecular markers.

The utility of whole-exome sequencing for genetic diagnosis of heterogeneous background disorders: an epilepsy and delayed psychomotor development case report

Maíra Cristina Freire, Michele Pereira, Giovana Torrezan, Mariano Zalis, Elvis Mateo, Alessandro Ferreira

Hermes Pardini Institute, Progenética - Hermes Pardini Institute

Abstract

Next-Generation Sequencing (NGS) technologies and more specifically the Whole Exome Sequencing (WES) has emerged as a successful tool for the diagnosis of genetic syndromes and has been particularly effective in identifying rare disease-associated genes. Furthermore, these technologies allow reductions in cost and time to diagnosis since it interrogates exome wide variation in a single assay and can provide a genetic assessment, even when candidate genes are not known or when a disorder exhibits substantial phenotypic and genetic heterogeneity. The identification of genetic bases of rare monogenic diseases is of utmost importance since provides important knowledge about risk prediction, disease mechanisms, biological pathways and potential therapeutic targets. In this context, clinical groups have shown the utility of WES in improving the diagnosis of rare genetic diseases. The present work describes a study of a family of consanguineous parents who had two daughters with the same clinical condition, delayed psychomotor development and epilepsy. The first daughter died without a diagnostic. Epilepsies and delayed psychomotor development has a highly heterogeneous background with a strong genetic contribution. Since many mutations in different genes are described for being associated with these conditions, we performed WES of the proband in order to identify the responsible gene. Blood DNA was extracted and exome capture was perfomed by Nextera Exome Capture System. Whole exome was sequenced on Illumina HiSeq 2500 platform and aligned to the hg19/GRCh37 reference genome, resulting in 95.9% of the target being covered at least at 10X with an average coverage of 126-fold. We identified a homozygous variant (c.637T>C/ Chr10:135179582) in exon 6 of the ECHS1 gene (NM_004092), which results in an exchange of cysteine to arginine on position 213 of the protein (p.Cys213Arg). This variant is located in a region well conserved throughout evolution of ECHS1 (PhyloP>2) and has not been reported in the dbSNP141, in the 1000 Genomes Project, in ESP6500 nor in eight thousand controls Brazilian and foreign. This mutation was predict to be disease causing (probability: 0.999) on Mutation Taster, damaging (score: 0) by SIFT and possibly damaging (score: 0.938) by PolyPhen-2. The ECHS1 is an important mitochondrial enzyme and its deficiency has been related with a new class of metabolic disease associated with mortality and morbidity rates in infants and younger children. This study reinforces the importance of the WES on diagnosis of rare genetic diseases with phenotypic similarity with others diseases.

OPTICAL MAPPING TO DETECT MISASSEMBLIES IN GENOME OF CORYNEBACTERIUM PSEUDOTUBERCULOSIS STRAIN 1002

Thiago Jesus Sousa, Diego Cesar Batista Mariano, Flavia Figueira Aburjaile, Flávia Souza Rocha, Felipe Luiz Pereira, Henrique Cesar Pereira Figueiredo, Artur Silva, Rommel Thiago Jucá Ramos, Vasco Azevedo

Federal University of Minas Gerais, Federal University of Pará

Abstract

Corynebacterium pseudotuberculosis is a bacterial species belongs to CMNR group, which includes species of the genera Corynebacterium, Mycobacterium, Nocardia, and Rhodococcus. This species is responsible for diseases that cause great economic losses in the whole world: Caseous Lymphadenitis (CLA) in sheep and goats (C. pseudotuberculosis biovar ovis); and Ulcerative Lymphangitis in horses (C. pseudotuberculosis biovar equi). The first assembly and annotation of C. pseudotuberculosis strain 1002 (Cp1002) were deposited at GenBank in August of 2010 in previous work of our group. The sequencing was realized using 454 Roche and Sanger technologies, and the assembly was made with Newbler software (Roche, USA). Close related strains was used as reference to produce a super scaffold and fill remaining gaps. Considering the advent of new sequencing and assembly strategies that provide higher quality data, we decided upgrade this genome, especially by the importance of an accurate genome to work in the others omics areas in our research group. A re-sequencing of Cp1002 was performed using the Ion Torrent 200 bp Sequencing kit on PGM benchtop machine, resulting in 731,481 reads with an expected coverage of $\sim$58-fold. Also, an optical map was acquired from OpGen Inc. (Gaithersburg, MD, USA), using the restriction enzyme KpnI. A new assembly was done using SIMBA software and 9 contigs was obtained, with a N50 value of 402,955 bp. An alignment between the optical map and the in silico restriction map of the contigs has shown that the assembly present a high completeness. Afterward the gap filling step was conducted by CLC Genomic Workbench (Qiagen, USA), using the super scaffold generated in the alignment. As result, the genome was comprised in a circular chromosome with length of 2,335,099 bp. At this work, we highlight an inversion with size larger than half length in genome of former assembly. The validation using WGM has shown as an effective strategy to detect misassemblies in genomes C. pseudotuberculosis strains, even when others strains were used to an assembly reference-based strategy.

EVOLUTIONARY AND FUNCTIONAL GENOMICS OF PSEUDOGENES IN TRYPANOSOMATIDS

Marcio silva, Marcos Catanho, Fernando Valín, Antonio Basilio Miranda

Instituto Oswaldo Cruz - Fiocruz, Facultad de Ciencias, Universidad de la República, Montevideo, Uruguay

Abstract

Trypanosomatids are protozoan parasites belonging to the order Kinetoplastida, an ancient group in the phylogenetic tree of eukaryotes. The sequencing of the genomes of some pathogenic trypanosomatids belonging to the genera Leishmania and Trypanosoma, has undoubtedly contributed to the understanding of the biology of these organisms as well as of relevant aspects of the evolution of their genomes. The current availability of several complete genomes of trypanosomatids, at various degrees of divergence, allows for robust comparative and evolutionary analyses, offering new opportunities to better understand important biological processes, or even reveal other, yet unknown, biological processes in these organisms. In this work, we aim to perform a comparative and evolutionary analysis of the processes and mechanisms related to pseudogenes in Trypanosoma and Leishmania species. We will try to identify genes or groups of genes which are prone to defunctionalization, and therefore to pseudogenization, as well as genes or groups of genes which have not been affected by these processes, and therefore are indispensable in these organisms. Such study will allow us to acquire a qualitative and quantitative overview of the processes of acquisition and loss of genes over time in these parasites. Another important aspect to be explored in this work is the possible acquisition of new functions by pseudogenes. The traditional view is that once function is lost, the fate of these sequences is the progressive degradation. However, there are considerable evidences that pseudogenes may acquire new functions, particularly involved in gene expression regulation of other members of the same multigene family. Thus, using deep sequencing data (Illumina), we will determine the expression patterns of pseudogenes in two trypanosomatids, T. cruzi and T. vivax, in order to get clues about their possible new functions. Hence, in this work, we aim to contribute to a better understanding of the dynamics of genomes of trypanosomatids, (i) obtaining a qualitative and quantitative overview of the processes of acquisition and loss of genes in these organisms over time, tracing the history of each particular gene family among the lineages studied, as well as (ii) analyzing the evolutionary dynamics of pseudogenes, i.e., the processes of pseudogenization and neofuncionalization in these parasites.
Financial Support CAPES, PAPES-FIOCRUZ, CNPq, FAPERJ, and Plataforma de Bioinformática Fiocruz RPT04-A/RJ

Identification of Non-homologous Isofunctional Enzymes (NISE) between Solanum lycopersicum and the phytopathogens Botritys cinerea and Fusarium Oxyosporum

Rangeline Silva, Leandro Pereira, Antonio Miranda

FIOCRUZ, Pontifícia Universidade Católica do Rio Grande do Sul

Abstract

Solanum lycopersicum is one of the most consumed crops worldwide and is subject to infection by various pathogens. Despite recent progress in sequencing and annotation procedures, the metabolic pathways of several organisms are not completely understood, which is unfortunate because these gaps can be the key to the understanding of many processes involving the mechanisms of infection and resistance. Since most genes are identified and annotated by their sequence similarity with previously annotated genes, it is likely that some of the many unknown enzymes in these processes are NISEs (Non-homologous Isofunctional Enzymes), also known as analogous enzymes. NISE display the same biochemical function but come from different evolutionary origins, being the result of convergent evolution. The aim of this study was to identify NISE between S. lycopersicum and its pathogens Botritys cinerea and Fusarium oxyosporum. Sequence data was downloaded from KEGG, release 73.1. Clustering and functional inference were obtained using the software AnEnPi. The identified NISES had their folds sorted using the SCOP and SUPERFAMILY databases. The cases of NISEs found between S. lycopersicum and B. cinerea belong to twelve different ECs (Enzyme Comission Number). Some have been validated following further analysis, including the proteasome subunit Y7 (EC 3.4.25.1), which plays an important role in the regulation of cell proliferation or cell cycle control, transcriptional regulation, immune and stress response, cell differentiation and apoptosis, basic mechanisms involved in the process of infection and disease. It also interacts with important proteins involved in cell cycle transition, transcription factor regulation, viral replication and even tumor initiation and progression. NISEs belonging to eight different ECs were found between S. lycopersicum and F. oxyosporum, such as the eukaryotic translation initiation factor 3 subunit H (EC 5.2.1.8), which is required for several steps in the initiation of protein synthesis and the serine/threonine-protein phosphatase 2A (PP2A) activator (EC 5.2.1.8), which acts as a regulatory subunit for PP2A modulating its activity or substrate specificity. Besides the identification of NISEs between the genomes of S. lycopersicum and B. cinerea/F. oxyosporum, our procedure revealed several enzymes that were not previously annotated in these genomes. Further work will improve not only genome annotation and NISE identification, but also our understanding of the importance of convergent evolution in shaping the genomes and a more detailed knowledge of the metabolism. Also, since some NISEs may be involved in host/pathogen interactions, a better comprehension of these processes may be achieved.

Comparative genomics of probiotic yeasts: Saccharomyces cerevisiae var. boulardii and S. cerevisiae UFMG A-905.

Thiago Mafra Batista, Rennan Garcias Moreira, Ieso de Miranda Castro, Carlos Augusto Rosa, Jacques Robert Nicoli, Gloria Regina Franco

UFMG, UFOP

Abstract

Probiotics are living microorganisms present in food and supplements that when ingested in sufficient amounts can confer health benefits. The yeast Saccharomyces cerevisiae var. boulardii was isolated by Henri Boulard in 1920 during a cholera outbreak in Indochina (current Vietnam). It is a non-pathogenic yeast, thermo-tolerant, and consists of the only eukaryotic microorganism marketed worldwide in the form of probiotics that is widely used in human medicine for gastrointestinal disorders treatment. The yeast Saccharomyces cerevisiae UFMG A-905 was isolated from a collection of S. cerevisiae that had been tested in vitro using simulated gastrointestinal conditions. It has also been tested in vivo for its ability to colonize mice gastrointestinal tract without causing any pathology. Its protective effect was demonstrated over gnotobiotic animals challenged with Salmonella typhimurium, Escherichia coli and Clostridium difficile. This characterizes, for the first time, a potential probiotic product of Brazilian origin. The use of strains of Saccharomyces as probiotics is advantageous because they do not exhibit resistance to antibiotics used in human therapy. In this study, we present the genome sequence of the UFMG A-905 and the three S. boulardii strains. All strains have characteristic genome sizes of S. cerevisiae strains, ranging from 11,4Mb to 11,6Mb and 5,350 predicted protein-coding genes, on average. The number of Gene Ontology terms, protein domains and Enzyme Code counts were similar among the probiotic strains. The phylogenetic relationships inferred from the alignment of 415 orthologous proteins and construction of a mega-tree by the Neighbor-Joining method were capable of clustering three major clades formed by strains of industrial importance, laboratory strains and alcoholic fermentation strains, where the probiotic strains were grouped. SNVs analyses suggest a conservation of variations in the probiotics genomes, with 51,943 average variants in each genome, a variant every 230 bases on average, and 70% impact of these variants among the four genomes. Twenty unique genes were found to be present in probiotic strains and absent in the non-probiotic strain S288c, most of them coding for proteins with unknown function. We found 803 S288c genes absent in the probiotic strains. From these, 706 were coding for proteins related to transposition activity and retrotransposons. The absence in the probiotic strains of some genes related to the influx of protons and ions, as well as heat shock proteins and chaperones, suggests an involvement of such genes in the cell resistance to acid stress.

Metabolic relationship of three strains of Lactococcus genre with an alternative carbon source

Tessália Diniz Luerce-Saraiva, Carlos Augusto Almeida Diniz, Sara Heloisa Silva, Rodrigo Dias Oliveira Carvalho, Cassiana Severino de Souza, Marcela de Azevedo, Fillipe Luiz Rosa do Carmo, Pamela Mancha Agresti, Mariana Martins Drumond, Izabela Ibraim, Letícia Castro, Siomar de Castro Soares, Henrique Figueiredo, Vasco Azevedo,

Federal University of Minas Gerais, Federal University of Triângulo Mineiro

Abstract

Lactic acid bacteria (LAB) are a heterogeneous group of Gram-positive bacteria, having as common the metabolic property to convert sugar into lactic acid as the main end product. LAB have long been used in food preservation and production. The majority LAB species are considered safe for human and animal consumption and many of them are also used as probiotic. The interaction between strains and gastrointestinal tract (GIT) is an important factor to evaluate the potential of probiotics during screening. Usually, their resistance to GIT and its adhesion to the epithelium has been evaluated as part of the selection criteria. Recently it was demonstrated that the metabolism of different carbon sources can influence the adhesiveness of the bacteria in GIT, in other words, alternative carbon sources can promote phenotypic characteristics in bacteria to perform its beneficial effect. In this work, we aimed to evidence the metabolic relationship of three strains of Lactococcus genre with an alternative carbon source by monitoring their growth in vitro and in silico analysis of their metabolic pathway integrity. For this, the strains used were Lactococcus lactis subsp. lactis NCDO2118 (LLNCDO2118 - a probiotic bacterium, which demonstrated anti-inflammatory activity in the treatment of chemically induced ulcerative colitis in a murine model), Lactococcus lactis subsp. lactis IL1403 (LLIL1403 - strain used extensively for the production of various metabolic products and production of recombinant protein), and Lactococcus lactis subsp. cremoris MG1363 (LLMG1363 – strain most commonly used in genetic and physiological research). These bacteria were cultured in complex medium supplemented with the preferred carbon source (glucose) or an alternative source (xylose) for the determination of growth curve. The three strains of Lactococcus had similar growth when cultured in medium supplemented with glucose, but the strains LLIL1403 and LLMG1363 exhibited limited growth when the carbon source was xylose. The analysis on the integrity of the metabolic pathway of xylose in these strains was determined using the Pathway Tools software (http://bioinformatics.ai.sri.com/ptools/). The metabolic limitation observed in LLIL1403 and LLMG1363 strains could be related to its steady growth in environments with nutrient-rich and its adaptation to these, which can cause loss of gene function and genome reduction, to have limited ability to live outside this environment. In other hand, LLNCDO2118 strain, an isolated from vegetables proved to keep this phenotypic characteristic. To confirm that the use of xylose by LLNCDO2118 strain will enhance its beneficial effect, in vivo experiments will be performed.

In Silico analysis of Single Nucleotide Polymorphisms (SNPs) in human FANCA gene

Abubaker Hamid Mohamed Salih, Ozaz Mohamed, Sami Salam, Hadeel Yousif, Mohamed Hassan

Africa City of Technology, University of Gezir University of Khartoum, University of Tuebingen

Abstract

Single-nucleotide polymorphisms (SNPs) play a major role in the understanding of the genetic basis of many complex human diseases. Also, the genetics of human phenotype variation could be understood by knowing the functions of these SNPs owing to the importance of FANCA gene in a post replication repair or a cell cycle checkpoint function. In this work, we have analyzed the genetic variation that can alter the expression and the function of the FANCA gene using computational methods. Genomic analysis of FANCA was initiated Polyphen and SIFT server used to retrieve 16 harmful mutations, among of these 16 nsSNPs damaged SNPs five non-synonymous SNPs showed very damaging by higher PSIC score of the Polyphen server with a SIFT tolerance index of 0.00-0.01 (R318M, I493T, A610T, P739L, R1117G). Protein structural analysis with these amino acid variants was performed by using I-Mutant and Modeling amino acid substitution with chimera software to check their stability and the effect of the native and mutant residues protein and structure for all 16 nsSNPs damaged. Screening for these SNPs variants in coding region may be useful for Fanconi anemia disease molecular diagnosis. Of the total 229 SNPs in 3'UTR region of FANCA gene, 24 SNPs were found in the 3' UTR contain alleles can be disrupts a conserved miRNA site, therefore might change the protein expression levels. In order to make effective use of genetic diagnosis, the harm SNPs in all fanconi genes should be well known and available to the diagnostic services and molecular biology laboratories to ensure accurate diagnosis for this complicated disease which can also lead to successful intervention which dependent on finding the cause or causes of a problem.

Genome assembly of Bothrops jararaca

George Willian Condomitti Epamino, João Carlos Setubal, Inácio Junqueira Azevedo, Milton Yutaka Nishiyama Junior, Diego Dantas Almeida

Universidade de São Paulo, Instituto Butantan,

Abstract

We present here preliminary results of the genome assembly of the serpent Bothrops jararaca. This study was motivated by the relatively few reptile genomes currently available and by the potential of novel toxin-coding genes discovery. The genome size is estimated to have 2.2 Gbp, in 36 chromosomes. Sequencing was done using a variety of plataforms, Illumina paired-end and PacBio, with a total of 6.723.545.770 reads. Even with this high coverage we have found out that the assembly of this genome is more challenging than expected, probably because of a high repeat content and high degree of heterozygosity. We have tested several assembly programs. Currently, the best results have been obtained with the following pipelines: pre-assemble reads into contigs with Abyss assembler, using paired-end reads; use the contigs produced and mate-pair reads as input for scaffolding using SSPACE; close the gaps in scaffolds using GapCloser; another set of good results have been obtained by running AllPaths_LG with paired-end and mate pair reads. The total wall time spent with the two pipelines was 1032 hours in a dedicated Linux server equipped with Intel I7 processor, capacity for 80 parallel threads and 1TB of RAM. We have run this same pipeline on publicly available reads from another serpent (Boa constrictor) as a control. The results for the control are much better than for B. jararaca, which is evidence for the difficulty in assembling this genome. All the results were evaluated with analysis tools like QUAST, which gives basic assembly metrics and CEGMA, which tries to find conserved eukaryotic genes in a given assembly.

Bioinformatics analysis of DNA alterations to identify mutations in evolved industrial yeast by evolutionary engeneering

Sheila T. Nagamatsu, Leandro V. dos Santos, Gonçalo A. G. Pereira, Marcelo F. Carazzolle

UNICAMP, UNICAMP/Biocelere

Abstract

Brazil is one of the world's largest first-generation ethanol producers behind only United States. The second-generation ethanol is a new and promising technology that can dramatically reduce the costs and increase the production. While the first-generation is based on fermentable sugar from sugarcane, the second, is based on hydrolyzed biomass consisting of the residual non-food crops such as leaves, stems, grass, etc. Both the first and second generations are dependent on the development of industrial yeasts which has a robustness phenotype to support the stress generated by industrial process. For the first-generation, industrial yeast should present a robust growth in order to be competitive with contaminant microorganism (wild yeast and bacteria), however, for the second-generation, is required robustness for some inhibitors (acetic acid, furfural and HMF) and non-fermenting sugar consumption (mainly xylose) obtained by biomass hydrolysis process. All these characteristics might not be founded in only one strain yeast, but, because the high variability of this specie, some traits can be identified in isolated strains such as sake yeast, wine yeast, ethanol yeast and wild yeast. Alternatively, the evolutionary engineering is an efficient methodology to improve some characteristic by several rounds of cell growth and recycling on selective growth media. In this context, the development of bioinformatics methodologies applied to yeast genome sequencing such as ab-initio genome assembly, comparative genomics, copy number variation (CNV) and SNPs/Indels identification are fundamental to study these natural or induced phenotypes identifying target genes for genetic engineering of industrial yeasts. In this study, transgenic xylose-consuming yeast (parental strain) was evolved by evolutionary engineering, using xylose as carbon source, to increase the xylose consumption ratio which resulted in five strains (three flocculated and two non-flocculated). Firstly, these six strains were sequenced and the parental genome was assembled to be used as reference. After that, reads generated from evolved strains were aligned into parental genome and submitted to SNPs/indels calling and annotation analysis using GATK and VEP software. After applying the quality filters, a total of twenty SNPs and five indels were identified and annotated. With this annotation we could select a set of interesting mutated genes for the five evolved strains, which only one was presented in all of the strains, two in flocculated strains and one in non-flocculated strains. Additionally, CNV analysis has been implemented to complement SNPs/indels information, generating more insights about the molecular mechanisms involved in the phenotype-genotype correlation.

Bacterial community profiling of human rectal cancers

Andrew Maltez Thomas, Eliane Camargo de Jeses, Ademar Lopes, Samuel Aguiar Junior, Ariana Ferrari João Carlos Setubal, Diana Noronha Nunes, Emmanuel Dias-Neto

Universidade de São Paulo, AC Camargo Cancer Center

Abstract

Colorectal cancer (CRC) is the third leading cause of cancer death worldwide and is responsible for more than 50,000 annual deaths in the United States alone. CRCs can be classified as inherited (genomically unstable), inflammatory (due to chronic inflammation of the gastrointestinal tract, e.g. Crohn's disease) or sporadic, the latter accounting for more than 80% of all cases. Recent publications have shown mechanistic evidence for the involvement of gut bacteria in the development of both inflammatory and sporadic CRC by means of genotoxin and DNA damaging superoxide radicals production, T-helper cell-dependent induction of cell proliferation, and Toll-like receptor mediated induction of pro-carcinogenic pathways. However, despite this vast body of circumstantial evidence, studies thus far have not been able to identify, in rectal cancer patients, key microbial species involved in such carcinogenic mechanisms. To aid in the description of species associated with sporadic rectal cancer, we compared bacterial communities of 18 rectal cancer patients and 18 individuals without rectal cancer by large-scale 16S rRNA metagenomic sequencing. Samples were collected after exploratory colonoscopy (non-cancer group) or surgery for tumor excision (rectal cancer group), DNA was extracted and the hypervariable regions V4-V5 of the 16S rRNA gene were PCR amplified and sequenced on the Ion PGM platform. We filtered sequences using Qiime and formed clusters of operational taxonomic units (OTUs) at 97% sequence identity using UPARSE. Representative sequences of each OTU were used for taxonomic classification using RDP classifier. We obtained a total of 5,593,020 filtered sequences with a mean length of 315 $\pm$ 30nt. After filtering to require OTUs with at least 3 sequences and present in $\geq$25% of all samples, 1,955 OTUs remained. To investigate differences in OTU, phyla and genera abundances between both groups, raw sequence counts were normalized then log transformed. Rectal cancer samples had higher log abundances of Bacteroides, Ruminococcus, Parabacteroides and Roseburia and non-cancer samples had higher log abundances of Pseudomonas, Paracocccus, Lactobacillus and Bacillus. Two OTUs assigned to Bacteroides fragilis were significantly more abundant in rectal cancer samples than non-cancer samples. Also in the Bacteroides genus, two other species, B. uniformis and B. ovatus, also had significant increases in rectal cancer samples. In non-cancer samples, Pseudomonas nitroreducens, Brevundimonas diminuta and Prevotella melaninogenica had higher log abundances than in tumor samples.

Identification and classification of microexon genes in a collection of genome annotations

Bruno Souza, Murilo Amaral, João Carlos Setubal, Sergio Verjovski-Almeida

Universidade de São Paulo, Instituto Butantan

Abstract

Microexon genes (MEGs) have an unusual architecture, composed predominantly by four or more very small tandemly disposed symmetric exons (microexons $\leq$ 36 bp, with exon sizes multiple of 3). They were first described in 2009 in the parasitic platyhelminth Schistosoma mansoni (etiologic agent of schistosomiasis). Some of those genes display evidence (by transcriptomics and proteomics) of generating variable protein isoforms through alternative splicing (employing a 'pick and mix' strategy). This, added to the fact that those proteins are secreted and their expression occurs mostly in the intra-mammalian stages of the parasite, led to the hypothesis that MEGs play a role in the escape mechanism from the host immunological system. The first described MEGs have no homologs outside the Schistosoma genus. MEGs have also been reported in other parasitic platyhelminths (Echinococcus granulosus and E. multilocularis), but they show no sequence similarity to those reported for S. mansoni. The presence of microexons has also been reported in genes of model organisms and humans, usually as a hotspot of alternative splicing, but in those cases only a single microexon per gene was observed. Within this background, we asked the following question: are there genes with similar architecture (i.e., with multiple internal microexons in tandem) in other organisms? We developed a heuristic to detect genes with the architecture of S. mansoni MEGs and applied it on a collection of genome annotations. This heuristic successfully detected all the original MEGs (including those reported on E. granulosus and E. multilocularis) and almost five hundred other genes distributed among 120 organisms (including animals, plants, fungi, and some unicellular eukaryotes). In this work we present details of these discoveries. Financial Support: CAPES, CNPq and FAPESP.

Missense mutations in candidate genes for reproductive disorders in a Gir bull identified through whole-genome sequencing

Ana Emília de Paiva, Pablo Augusto de Souza Fonseca, Fernanda Caroline dos Santos, Izinara Rosse da Cruz, Guilherme Silva Moura

UFMG

Abstract

Fertility of dairy cattle has decreased in the last decades. Most of reproductive failure in cattle is caused by gonadal hypoplasia associated or not with spermatic abnormalities and many studies have tried to understand its genetic causes. A widely used approach for the investigation of genetic causes of complex traits is the Genome-Wide Association Study (GWAS). However, the candidate marker identified through GWAS may be just reflecting a linkage disequilibrium relationship with the real causal variant. Therefore, further studies in candidate regions must be conducted in order to ascertain the real causal variants. In a previous GWAS, performed by our research group, three regions on the X-chromosome associated with gonadal hypoplasia and spermatic abnormalities in Dairy-Gir were identified. In order to identify the causal variants for these phenotypes, a whole-genome sequencing of the individual presenting the highest number of spermatic and testicular defects was performed. Sequences were obtained using the Illumina Miseq platform. Duplicated reads, and those with Phred<30 were removed using the softwares Picard and Trimmomatic, respectively. Quality control was performed using the Fastqc software. The remaining reads were mapped against the UMD 3.1 reference of the bovine genome. The variations were identified using the Samtools software and were subjected to a quality control filtering in order to ensure the reliability of them. The variations identified in the affected individual sequences were compared with the variations of three healthy Gir and only the variations present exclusively at the affected bull were used in the further analyses. Functional annotation of them was performed through the NGS-SNP software. Annotated SNPs located on the three X-chromosome regions previously identified were selected. Among this set of SNPs we were able to identify missense mutations at genes that plays important roles in the control of the cellular cycle and apoptosis. Genotyping and association tests will be performed in a larger sample in order to investigate the potential effect of these mutations for spermatogenesis and bull fertility. This is the first study to investigate causal variants for reproductive disorders in Dairy-Gir. Results obtained in the next steps may help to improve genetic breeding programs and to understand the metabolic and physiological processes involved in male infertility in Dairy-Gir and others breeds.

Using the Mean Shift clustering algorithm to predict Genomic Islands in bacteria

Daniel Miranda de Brito, Thaís De Almeida Ratis Ramos, Vinicius Maracaja-Coutinho, Sávio Torres de Farias, Leonardo Vidal Batista, Thaís Gaudencio do Rêgo

Departamento de Informática, Centro de Informática, Universidade Federal da Paraíba
Universidad Mayor

Abstract

In recent years, the cost reduction in genome sequencing was responsible for a considerable increase in the number of bacteria with its whole genome sequenced. The availability of these large volumes of genomic sequences has brought important information related to the genomic structure to be explored, especially for Genomic Islands (GIs). GIs are regions in bacterial genomes that were acquired from other organisms by the mechanism of horizontal gene transfer (HGT). These regions are often responsible for many important adaptations to the bacteria, with great impact on its evolution and behavior. Nonetheless, these adaptations may negatively impact human and other species health and are usually related to pathogenicity, drug resistance and virulence. The identification of these regions allows researchers to identify genes responsible for these conditions and develop new vaccines and antibiotics. For this reason, many computational approaches have been presented for GI prediction, however, its efficacy still requires improvement. We develop a new method to predict GIs in bacteria, built upon Mean Shift clustering algorithm, that does not require the definition of the k number of clusters for the prediction, but requires the bandwidth parameter. For that, a heuristic to determine this value automatically from the selection of artificial islands genome inserted in the genome was developed. The method is based on sequence composition, i.e., it uses the fact that genes of a particular specie are normally similar enough on its base composition, thus sequences acquired from other organisms can be distinguished by sequence analysis. In the proposed method, we use the base vector <A, T, C, G> to capture the genomic signature. We applied the clustering algorithm in disjoint regions from the genome, in order to separate the potential horizontally acquired genes from the original genes from the host genome, allowing its identification. The method has been tested in bacteria with known genomic islands, discussed in other papers, and its application revealed the same GIs predicted by other methods and novel regions, not yet predicted. Detailed investigation in the new predicted islands found the presence of typical GIs elements, confirming its effectiveness.

Analysis of fecal bacterial diversity in howler monkeys (Alouatta) through metagenomics

Raquel Franco, Arthur Berselli, Layla Martins, Andrew Thomaz, João Batista, Julio de Oliveira, Aline Silva, João Setubal

USP, FPZSP, UNIFESP

Abstract

Many vertebrates depend on vegetable matter-based diets as the main energy source. In primates, as in other vertebrates, digestion occurs along the gastrointestinal (GI) tract assisted by symbiotic microorganisms, which increases the host digestive efficiency. According to the literature, GI microbiota composition and diversity are strongly affected by the host diet and habitat. In howler monkeys, previous studies revealed divergence in the GI microbiota between individuals that inhabit different regions in southeastern Mexico in both wild and captivity. Some species of howler monkeys (e.g. Alouatta guariba clamitans and Alouatta caraya) are considered threatened and vulnerable, and reintroduction of animals into the wild, after captive breeding, is a viable alternative for species preservation. Investigation of the impact of dietary habits in the GI microbiota of howler monkeys species may contribute to a better adaptation of these animals in captivity and in the wild. It has been reported that in the wild these species of howler monkeys feed mainly on leaves (65%), fruits (25%) and flowers (10%), while captive individuals are generally fed with animal food mixture (2%), fruits (58%) and leaves (40%). In this project, we aim to investigate the GI microbiota composition and diversity of A. guariba clamitans and A. caraya by 16S rRNA amplicon sequencing of fecal DNA samples collected from noncaptive and captive individuals that inhabit Sao Paulo Zoo Park. Noncaptive animals are those that live freely in the Park Rain Forest patch. Fecal samples from three captive and three noncaptive animals were collected in Spring and Autumn, total DNA was extracted and used for PCR reactions with customized primer pair for amplification of the variable region V3 and V4 of 16S rRNA gene. The 550 bp-amplicon libraries were subjected to sequencing using Illumina/MiSeq Reagent kit v2 (500-cycle format, paired-end (PE) reads) and 6 bp-overlapping PE reads were assembled using the Fastq-join Aronesty software. After quality-filtering and trimming, amplicon sequences were clustered, chimera-filtered and mapped to OTUs (Operational Taxonomic Unities) database for abundance estimation, using UPARSE. Taxonomic identification and diversity analyses were performed with QIIME. Initial results revealed that the abundance of different taxa varies between captive and noncaptive individuals. Details of these differences will be presented. Supported by FAPESP and CAPES.

Computational methods for metagenomic data processing in the Metazoo Project

Gianluca Major, Felipe Lima, Andrew Thomas, Leandro Lemos, Deyvid Amgarten, Deibs Barbosa, Calos Morais, Luciana Antunes, Aline Silva, João Setubal

USP

Abstract

The Metazoo project aims to study the microbial communities in three different environments in the São Paulos Zoo Park: a composting process, the Sao Francisco Lake, and feces of resident howler monkeys, by using a metagenomics approach. Different computational methodologies are being used to analyze the sequences datasets that were generated from sequencing total DNA, 16S rRNA amplicon and mRNA (RNA-seq) with Roche-454 and/or Illumina (MiSeq and HiSeq2500). (all three kinds) platforms. The sequences derived from time-series samples for both the Lake and the composting or from six howler individuals. For each time point, depending on the type of data, the following steps were taken: (i) total DNA shotgun sequences were analyzed using MyTaxa in order to obtain a taxonomic classification, assembled using SOAPdenovo to generate contigs which were then submitted to the IMG/M platform to obtain a functional identification using annotation terms such as COG, KO and EC; (ii) mRNA data was assembled into contigs using SOAPdenovo (metatranscriptome version) and submitted to the IMG/M platform, using the same annotation process; predicted coding sequences were then analyzed using MyTaxa for taxonomic classification; (iii) amplicon 16S rRNA sequencing data were clustered into OTUs using 97% sequence identity and the UPARSE program, and representative sequences were taxonomically classified using the RDP classifier and Qiime. All assemblies were preceded by a quality filtering step where sequences were filtered by their mean quality score using SICKLE. The taxonomic analyses were complemented with an interactive visualization of taxon abundances using KRONA. We also analyzed the metabolic potential, via KEGG pathways, of total DNA sequencing data using MEGAN and HUMAnN. The abundance data from both the taxonomic and functional analyses allowed us to analyze the microbial community profile for each sample, allowing ustime point a better understanding of their dynamics and functional potentials, and at the same time identify expressed geneswere . To check the coherence between total DNA and mRNA sequencing data, we aligned mRNA sequences to total DNA sequences using BLAST. Besides analyzing the Metazoo projects datasets using computational methods, we are creating new computational tools, such as a metagenome browser and a conceptual framework and respective database to ease the integration of different types of data and metadata generated by metagenomic projects. Supported by FAPESP, CNPq and CAPES.

Metagenomic Analysis of Rumen from Cattle Fed With Different Levels of Mate Extract (Ilex Paraguariensis A.St.-Hil.)

Maurício Mudadu, Maurício Cantão, Larissa Gonçalves, Léa Chapaval, Teresa Alves, Wilson Malagó, Fabrício Correa, Marcio Rabelo, Daniel Cardoso

EMBRAPA, UNICEP, Universidade de São Paulo

Abstract

The rumen provides to cattle the ability to absorb and digest plant fibers like cellulose, xylan and complex carbohydrate, that are going to be efficiently converted into cattle's body mass. There are in the rumen of a bovine thousands of microorganisms, such as Bacteria (most diverse, 1,011 cells/mL of rumen fluid), Protozoa, Archea and Fungi. Is it possible to change the population of microorganisms of the rumen by feeding cattle with different dosages of Mate extract (ME), a herb well known and consumed by humans? To answer this question, we used different concentrations of this herb in the forage used to feed four groups of twelve Nelore steers: T1 (no ME), T2 (0.5% ME), T3 (1.0% ME) and T4 (1.5% ME). Rumen fluid were sampled into pools relative to each group. Total DNA from the pools was extracted and used to perform a whole-DNA sequencing using a MiSEQ Illumina machine (Pair ended, 2x300 base pairs). We used a pipeline for metagenome analysis that comprised the softwares FASTQC and seqyclean (quality control), Diamond (alignment of reads against public databases), MEGAN (taxonomic and functional analysis) and STAMP (differential analysis and figures). For a preliminary result, only groups T3 (1.0% ME) and T4 (1.5% ME) were tested. A quality control filtering resulted, for the groups T3 and T4 respectively, 4,139,997 and 2,894,142 paired reads (min. length 80bp, PHRED 20). Taxonomic profiling showed a high prevalence of the taxa Bacterioidetes (Phylum) and Prevotella (Genus) which are also more abundant in metagenome profiles of rumen found in literature, indicating that we had a good quality sequencing result. Functional analysis showed high proportion of sequences related to "Metabolism" (KEGG), "Carbohidrates", "Metabolism of proteins", "Aminoacids and derivatives" and "Respiration" (SEED), which also are up-regulated pathways found in rumen. All these high sampled taxa and functional groups cited are augmented in the T4 group in relation to the T3 group, indicating that ME seems to modify the population of microorganism that participate in the metabolism, growth and feed efficiency in cattle. More comparisons and sequences are needed in order to obtain a better statistical reliability for the results. Supported by FAPESP (2011/51555-7) and EMBRAPA.

METHOD FOR IDENTIFYING BCR-ABL1-LIKE PEDIATRIC ACUTE LYMPHOBLASTIC LEUKEMIA

Gabriel Lopes Centoducatte, André Bortolini Silveira, Silvia Regina Brandalise, José Andrés Yunes

Centro Infantil Boldrini

Abstract

Acute Lymphoblastic Leukemia (ALL) is the most common pediatric cancer type, corresponding to about 25% of all cases of cancers and 80% of all leukemias in patients of less than 15 years old. Despite the improvement on the prognosis for most ALL therapies, this disease is still the most common cause of cancer-related death in young people around the world. The ALL is a heterogeneous disease in terms of both genetic and clinical outcome and some molecular subtypes of ALL were shown to benefit from incorporation of new drugs or treatments intensification – a high risk t(9;22) chromosomal translocation (BCR-ABL1) positive ALL, for example, benefits from the use of imatinib. Ten percent of pediatric ALL, even without the t(9; 22) exhibit a very similar gene expression profile compared to the BCR-ABL1 ALL; this subtype was named B̈CR-ABL1-likeÄLL. Once the BCR-ABL1-like ALL does not have a specific-easy detectable molecular marker such as the 9:22 translocation, the objective of this study is to develop a fast and effective method for the identification of BCR-ABL1-like ALL at diagnosis by studying differences on the gene expression profiles among different ALL subtypes. We performed global gene expression profiling (Affymetrix Human Gene 1.0ST) of 117 patients from our institution with both precursor B and T-cell ALL. We studied the differences on the gene expression profile among all the cases to be able to identify a minimum set of genes whose differential expression could be used as a classifier of BCR-ABL1-like ALL and compare these results with published data from other locations. A gene expression signature for BCR-ABL1-like was evaluated and validated on independent cohorts from Boldrini Children Hospital and St Jude Hospital; this signature was used to design probes in order to identify the BCR-ABL1-like ALL via qPCR reaction and validate the differences of gene expression among the ALL cases at Boldrini. In this study we were able to identify a list of genes that can be used as potential genetic markers for the identification of the BCR-ABL1-like ALL at diagnosis and help to understand the biology of this ALL subtype.

State of the art of computational techniques applied for characterization and prediction of Transcription Factor Binding Sites (TFBS)

Antonio Ferrão Neto, Luiz Paulo Moura Andrioli, Ariane Machado Lima

University of São Paulo

Abstract

There are several possible approaches used to predict Transcription Factors Binding Sites (TFBS). We present here a systematic review of articles that discuss the prediction of TFBS, pointing out the characteristics of each approach. To the best of our knowledge there is no systematic review in this topic. Although most works are based on P̈osition Weight Matrices(P̈MWS), it was found a growing interest in using other ways of TFBS characterization. We observed that the incorporation of additional information to the prediction system, in different approaches, significantly improves the predictive capacity. Examples are the consideration of the interdependence between the DNA bases, the interactions between DNA molecules and transcription factors considering the three-dimensional molecular structure and addition of biological, biochemical, computational and statistical information to the system. There are computational and financial costs for incorporation of such information, in the case of a computer program. However, these costs have fallen in recent years, increasingly enabling the adoption of these new methods.These costs can also refer to the experimental effort required to get ïn vivoör ïn vitrodata by the use of molecular biology techniques. Biophysical or Biochemical techniques such as the X-ray diffraction crystallography or nuclear magnetic resonance (NMR), used for obtaining the structure of a protein and the study of the molecular docking between the homeodomains of proteins and TFBS in the DNA, or even every effort to perform a statistical analysis, etc. Some authors tried to combine some of these approaches, increasing the prediction performance, but there is no perfect solution and the prediction of TFBS with low false positive rate remains a challenge not yet solved.

Understanding the evolution of pathogenicity using comparative genomics of the fungus responsible for the wilt disease in cacao

Paulo Massanari Tokimatu, Juliana Jose, Gonçalo Amarante Guimarães Pereira, Leandro Costa Nacimento, Eddy Patricia Lopez Molano, Odalys Garcia Cabrera, Karina Yanagui, Marcelo Falsarella Carazzolle

Instituto de Biologia/UNICAMP

Abstract

The ascomycete fungus Ceratocystis cacaofunesta is responsible for the wilt disease in Theobroma cacao, causing significant losses in the production of cacao in South America. The fungus enters the plant through wounds and moves to the secondary xylem, causing necrosis of related parenchyma cells. The disease promotes the vessel occlusion, hence the symptoms of chlorosis and wilt of leaves, killing the host in two weeks after infection. The Ceratocystis genus is composed of non-pathogenic and pathogenic species with a large host range. There are five public genomes sequenced available for the genus and an outgroup of the same family, but they are not annotated for protein coding genes. Additionally, our group sequenced the C. cacaofunesta genome and transcriptome, generating an interesting dataset, allowing us to perform genome prediction, annotation and comparative analysis on these genomes. This work uses genomes of the Ceratocystis genus and RNA-seq data of in vitro C. cacaofunesta to predict genes and understand how gene families expanded and retracted in the pathogenic species C. cacaofunesta. We mapped the RNA-seq data to the genome of C. cacaofunesta using STAR. Knowing the intron locations, we created a training group and performed a gene prediction using Genemark-ET. Then, we filtered the results to find the most reliable putative genes using BlastP against sequenced proteins of the genus and parameters like small consecutive gaps. This group of reliable genes was used as a training group to generate hidden Markov models for another prediction tool, Augustus, and finally compared and combined the two predictions in EVM. Genemark and Augustus predicted 7910 and 7174 genes for C. cacaofunesta, respectively. The comparison between them showed that Genemark and Augustus found 992 and 253 unique genes, respectively. The same methodology was applied for the next 5 genomes generating a dataset that allow performing comparative genome approaches by assignment of orthologs and paralogs and estimating gene family expansion along species. Pathogenic and non-pathogenic species presented many different genomic characteristics as reduced gene sizes for the former and the expansion of families functionally different among them. Future steps will involve a detailed study of the genome, looking for groups of genes related to pathogenicity and resistance of C. cacaofunesta to the plant defense not only to search for targets to develop and improve countermeasures to the wilt diseases in cacao and other hosts, but also to understand how the pathogenicity evolved in the group.

Hybrid genome assembly of bacteria Burkholderia sacchari, a natural bioplastic producer

Pedro Nepomuceno, Paulo Alexandrino, José Gomez, Luiziana Silva, André Fujita

University of São Paulo

Abstract

Burkholderia sacchari is a natural biopolymer producer. It produces polyhydroxyalkanoates (PHAs) as storage material for energy reserve. However, the industrial use of this biopolymer is still not economically viable, though the analysis of metabolic fluxes may unearth a way of making it feasible. In this context, the knowledge about the genomic content of the bacteria is fundamental. Next Generation Sequencing technologies have made genome assembly easier due to its price, speed and massive amount of data produced in a short period of time. Although it is now possible to obtain draft genomes much faster, the assembly of contigs has become a much more complex and error vulnerable process than old-fashioned Sanger electrophoresis sequencing. There are now a various numbers of ways to assemble a genome, and also a huge number of ways to evaluate the results. The majority of metrics for this evaluation are based on the number of contigs, their average length, corrected N50 values, and number of annotated genes. Studies like GAGE-B and tools like QUAST come in handy when it comes to compare the algorithms. The Burkholderia sacchari genome was sequenced in 2012 using Roche/454 technology and the resulting reads have already been assembled and published. Furthermore, the bacteria was recently sequenced again using the Illumina technology. The aim of the present work is to assemble the hybrid genome using different strategies and evaluate these assemblies. The softwares Velvet and MIRA were used to assemble contigs with different sets of parameters and quality control techniques. Resulting assembles were compared using aforementioned comparison tools to assess the impact caused by parameters variation used over the metrics.

Reconstructing the Whole Mitochondrial DNA (mtDNA) from Nuclear Genome

Giovanni Marques de Castro, Adhemar Zerlotini Neto, Michel Eduardo Beleza Yamagishi

Embrapa Informática Agropecuária

Abstract

In several eukaryotic organisms, the nuclear genome has several partial copies of the mitochondrial DNA (mtDNA). These copies are called NUMTs (NUclear MiTochondrial DNA) and they have been known since 1967 when the first evidence of them were reported in the mouse nuclear genome. Despite almost fifty years have passed, the reason of their very existence remains controversial. However, their presence has been confirmed in an increasing number of genomes. The NUMts could be only another DNA idiosyncrasy, but they actually represent a serious issue for important application such as genome bar coding. There are many open questions about them. A practical one could be stated as: is it possible to reconstruct the whole mtDNA from the NUMTs, and how similar this hypothetical sequence would be to the actual mtDNA. In order to address this question, we have taken advantage of PacBio reads derived from a Bos indicus mtDNA-free sperm sample, and after a mapping procedure that sorted out nuclear from NUMT-like reads, the Mira assembler was applied to get a sequence that has 97% of identity to the actual Bos indicus mtDNA. Although our bovine experiment can not be straightforward extrapolated to other organisms without further investigations, it has helped us to answer the aforementioned question, and might suggest that DNA samples, where the mtDNA is lacking or so damaged to the point of preventing its assembly, can, nevertheless, deliver enough information to reconstruct a sequence that resembles the actual mtDNA. For instance, ancient DNA derived from fossils may be an example of interest.

In silico characterization and mapping of alpha-pheromone of P. lutzii

Juliana Alves Vieira, Waldeyr Mendes Cordeiro Silva, Maria Emília Machado Telles Walter, Ildinete Silva Pereira

Instituto Federal de Goiás (IFG), UnB

Abstract

The genus Paracoccidioides includes the two thermodimorphic species P. brasiliensis and P. lutzii, that occur in mycelial form at environmental temperatures (below 25^{o}C), and switch to a pathogenic multiple-budding yeast-form at the mammalian host temperature (37^{o}C). Both of which are the etiologic agents of paracoccidioidomycosis, a systemic mycosis that affects humans in Latin America, particularly in Brazil. For many years, Paracoccidioides was considered as an asexual and clonal microorganism, but recent evidence suggests that Paracoccidioides species have the potential to undergo sexual reproduction. Comparative genomic analysis of all dimorphic fungi and Saccharomyces cerevisiae demonstrated the presence of conserved genes involved in sexual reproduction, including those encoding mating regulators such as MAT, pheromone receptors, pheromone-processing enzymes, and mating signaling regulators. It is known that the alpha- pheromone has 49 amino-acid residues and is cleaved at the site KR by Kex2 peptidase, leaving 9 amino-acid residues in the mature alpha-pheromone. In this work, we identify the alpha-pheromone on the genome (from BROAD Institute - http://www.broadinstitute.org) and transcriptome (ESTM51 from NCBI by the GenBank Acc: BE758605) of P. lutzii. The deduced amino-acid sequences of all the ESTM51 ORFs were analyzed and, after confirmation of the presence of the alpha-pheromone sequence, the corresponding transcript was aligned against the genome of P. lutzii to perform the alpha-pheromone mapping. We propose an annotation of this gene on the genome with the identified exons, introns and CDS. In absence of a 3D model made by homology, we also propose a construction of such model using ab initio methods.

Genes implicated in cancer encompass both ancient and very recently originated molecular functions

Fernanda Stussi, Carlos Gonçalves, Miguel Ortega

Universidade Federal de Minas Gerais

Abstract

The origin of novel genes is great focus of interest. Preimplantation embryo development, for instance, is controlled by a more ancient gene, nanog (originated and shared in the coelomate clade) and a more recent one, Oct4 (originated and shared in the euteleostomi clade). However, both are transcription factors, a molecular function that is very ancient, since it is present in all cellular organisms. Interested on the first occurrence of a given molecular function along evolution, we set up to investigate this for all clades of the human lineage using Gene Ontology (GO). We began by determining the lowest common ancestor (LCA) for every term on the GO database, by checking the taxonomy IDs of the proteins annotated to them, and then we made some refinements to this LCA attribution. We attributed to every parent of a leaf term the child's LCA if it was more ancient than the parent's LCA. Next, we filtered out more general terms by deleting from protein annotations all terms with children (thus, more specific) terms annotated to the same protein. With this we have determined the Last Molecular Funcion (LMF) attained by every protein. To evaluate whether or not an LMF describes the gene appropriately, we analyzed genes implicated in cancer in Kegg Pathways. A total of 85 UniProt entries were curated. Most of them (89%) showed a great agreement between protein UniProt description, such as "Fibroblast growth factor receptor 3" and "fibroblast growth factor-activated receptor activity", while some demand additional annotation in GO database to add specification to the terms already annotated to them, e.g. Von Hippel-Lindau disease tumor suppressor (UniprotID: P40337), whose LMF is transcription factor binding, however it is shown to have the activity of ubiquitination of ADRB2; or Basket ProteinGTPase Hras (UniprotID: P01112) which is annotated as protein C-terminus binding but might be described as having the MAP kinase activation activity. Proteins that might receive more descriptive annotation comprised 11% of the analyzed ones. Moreover, 22 of these genes implicated in cancer pathways acquired the LMF by the Amniota clade, in comparison to 14 and 19 proteins that show Molecular Function which appeared in more ancient clades, Eukaryota and Cellular Organisms, respectively. Thus, genes implicated in cancer might encompass both ancient and very recently originated molecular functions.

Combined genomics, transcriptomics and proteomics strategies to improve annotation of transcribed and untranslated pseudogenes in Francisella noatunensis subsp. orientalis

Felipe Pereira, Guilherme C Tavares, Siomar C Soares, Alex F. Carvalho, Frederico A. A. Costa, Fernanda A. Dorela, Vasco A. C. Azevedo, Carlos A. G. Leal, Henrique C. P. Figueiredo

Aquacen/UFMG

Abstract

Francisella noatunensis subsp. orientalis (FNO) is an emerging pathogen that affects Nile tilapia farms around the world. FNO is a facultative intracellular pathogen and its genome seems to be undergoing a reductive evolution, containing a high number of pseudogenes (over three hundred per genome). However, there are no studies that validate the occurrence of this large number of pseudogenes through transcriptomics and proteomics analysis. The FNO strain FNO12, isolated in a franciselosis outbreak in Brazil, was selected for this study. The complete genome of this strain was available from previous work, and it was assembled using Nextera v3 MiSEQ dataset with 1,380-fold coverage. In silico prediction and annotation infer 363 pseudogenes to this strain. To validate this number, this bacterium was cultured in CHAH agar at 28o C for 96 h and then used to inoculate MMH broth equilibrated at 28o C for 24 h in triplicates. Thereafter, two aliquots of each replicate were collected for 2 trials. In the first trial we evaluated whole-genome transcriptomic through microarray-based gene expression analyses with a custom-made Agilent slide formulated based on the complete genome sequence. The second trial evaluated the protein expression by label-free shotgun proteomics using a Synapt G2Si mass spectrometer coupled to ion mobility. Transcriptomic and proteomic results were compared to verify if the pseudogenes were transcribed and further translated. Transcription of 325 of these pseudogenes were identified by microarray analysis, while only 8 related proteins were further translated in proteomic study. Thus, we verified that almost all pseudogenes annotated are really nonfunctional genes. The result support the in silico genome analysis, where these 8 pseudogenes could have been wrongly annotated by verifying the sequences already deposited. This result also corroborates the fact that the reductive evolution occurs in FNO and that matches with intracellular lifestyle of this pathogen in the fish host. The results obtained could validate the number of the pseudogenes in FNO12 strain through combination of transcriptomics and proteomics, allowing the improvement of genome annotation.

Quantifying Heritability of Copy Number Variation for Genome Wide Association Studies

Ana Claudia Ciconelle, Júlia Maria Pavan Soler

USP

Abstract

Copy-Number Variation (CNV) is an alteration in the number of copies of one or more sections of the DNA. These variation patterns along the genome can be sporadic or inherited and may be associated with various diseases or traits. The first motivation is to quantify heritability of CNVs considering dataset from Baependi Family Heart Studywith 119 Brazilian families (1,700 individuals). For CNV estimation Raw CEL files from scanned Affymetrix 6.0 arrays were processed using Affimetrix power tools (APT) and PennCNV to derive estimates of the relative copy number (log R ratios) and B allele frequencies(BAF) at each molecular marker.Using hidden Markov chain tools it was built a set of CNV regions, defined by the CNV regions overlaps of all samples. By analyzing the consensual CNVs along the genome and using a kinship matrix from the family structure, we identify CNVs with heritability of 30% or more. As a second motivation for genome wide association studies, the height phenotype, expected to have around 80% of heritability, was analyzed to describe the heritability and non-inherited components related to CNVs. GWAS (Genome-Wide Association Studies) succeeded to identify around 50 variants that may be associated with height, but they can explain only 5% of height variation in the population. Further, based on SNPs platform, it was found groups of polymorphisms that can describe up to 4% of the phenotype variation. Understanding the genetic model for height is a big challenge and through CNV analysis our results contribute for characterization of the pattern of missing heritability for this phenotype in the Brazilian population.We are applying the CNVs analysis for other phenotypes associated with cardiovascular diseases, such as blood pressure and glucose.

The use of Machine Learning to prediction of psychiatric disorders in children

Walkiria Resende, Henrique Cursino Vieira, Ana Cecília Feio, Fabricio Martins Lopes, Helena Brentani

USP, UFTPR

Abstract

Psychiatric disorders are the result of the interaction of several genes and the environment. Also during childhood, diagnosing of this disorders is not a trivial task, since different disorders share the same symptoms. Despite this difficulty, it is very important that it be done as soon as possible, which can ensure a better response to intervention. Besides, profile polygenic and multifactorial of these disorders reaffirms the difficulty of diagnosis. One gene may be associated with various disorders and different disorders share several genes, in other words genes do not encoding disease but behavioral endophenotypes. GWAS studies has contributed with polygenic scores for classification and risk prediction. However, some points are open as: necessity of studies with very large sample size; problems related to investigate populations of different genetic background; implementation of these genetic markers in clinical practice without any understanding of their biological significance and; non-inclusion of environmental component. So, to the Brazilian individuals, the use of European or American GWAS scores can introduce some bias because the population is mixed, having a different genetic component of other peoples of the world. Given the assumptions we are developing a prediction method that given a population case and control, knowing that there is genetic load and extrauterino and intrauterine environmental risk factors, make a prediction, classifying the samples of the population in healthy or ill to psychiatric disorders from a set of genes associated with human behavior. To do this we are using 1671 samples from INPD project, with 468 cases and 1203 controls, from Brazil. We genotyping 123 SNPs and ensure the consistency of the results according to analysis of HWE. This step is very important, since these errors occur very frequently and may impact the distribution of genotypes. With these analyzes we excluded 51 SNPs. We developed also a pre-processing methodology, consisting of data encoding, balancing and imputation of missing genotypes. Then we propose a collaboration to prediction at risk of childhood disorders, we will develop a classifier high sensitivity and specificity. To prevent insignificant representations or still a overfitting will be a selection of SNPs based on a priori knowledge about associations between genes and phenotypic dimensions. Dimensions such as aggression, impulsiveness, sociability, executive, functions present in different psychiatric disorders. The next challenge to be solved is the parameterization of the classifier, in which we propose to use bio-inspired heuristics. Finally we will train and validate the classifier.

A broad overview of somatic variants in childhood leukemia: Prospection and validation of collaborative mutation with the mutant IL7r in genome-scale data

Gisele Rodrigues, Lívia Campos, Priscila Zenatti, Elda Noronha, Maria Pombo-de-Oliveira, Silvia Brandalise, Francisco Lobo, José Yunes

Centro Infantil Boldrini, Instituto Nacional do Câncer

Abstract

The IL7/IL7R mediated signaling is essential for normal development and homeostasis of T cell precursors. Approximately 10% of patients with Acute lymphoblastic leukemia T cell (T-ALL) possess mutations in the alpha chain of the receptor for IL7 (IL-7Rα). Such mutations usually occur in exon 6 and in most cases introduce an unpaired cysteine in the extracellular juxtamembrane-transmembrane region. This mutation promotes de novo formation of intermolecular disulfide bonds between mutant IL-7Rα subunits, thereby driving constitutive signaling via JAK1 and independently of IL-7, gamma-chain or JAK3. Mutations such as the ones described for IL7R are important factors to initiate leukemia, but in many cases these changes alone are insufficient to achieve a complete leukemic phenotype, suggesting the occurrence of collaborative oncogenic mutations. To identify putative somatic mutations that may work in collaboration with the oncogenic mutated IL7R, we performed whole-exome sequencing and SNP-CNV-Array assays on a group of seven primary T-ALL samples carrying the IL7R mutation (T-ALL-IL7Rmut) and their paired remission samples (samples from the same patient after chemotherapy). We performed CNVs detection using ChAs (Chromosome Analysis Suite, version 2.0.1.2.). For the exome sequencing we used Illumina Hiseq2000 plataform and Agilent SureSelect V4 51M Capture kit. We performed read alignment in human genome reference genome version hg19 using bwa, followed by detection of somatic single nucleotide variations (SNVs) using Varscan2 (defined as mutations that occur only in leukemia samples when compared against paired remission samples). We performed detection of small indels using a similar strategy. We obtained a mean sequencing coverage of approximately 80X and 50X for leukemia and remission exome samples, respectively. We performed automatic variant detection followed by human curation and grouped all classes of somatic mutations surveyed in this study (SNVs, indels and CNVs) by genes, finding 17 of them to be recurrently mutated in IL7R mutated samples (i.e., mutated in at least two patients). After a literature survey we selected five of such genes for in vitro assays due to their previous involvement with ALL. To determine whether these recurrent mutations may collaborate with mutIL7R, BaF3 cells expressing or not mutant IL7R were transduced with pLKO.1 MISSION shRNA lentiviral vectors against the candidate genes or scramble controls. Preliminary results showed that all five candidate genes, when silenced, resulted in increased clonogenicity and proliferation of BaF3 cells in an IL7Rmut-dependent manner, suggesting they are possible suppressors genes in the IL7R-driven leukemogenesis. This work was supported by FAPESP.

Semi-Markov Conditional Random Fields for Gene Prediction

Ígor Bonadio, Alan Durham

USP

Abstract

Gene prediction is an open problem due the low accuracy in predicting correctly the gene structure. We can highlight two distinct approaches to solve this problem: (i) extrinsic, which is based on analysis of similarity between sequences like proteins or DNA stored in a database; (ii) intrinsic, which uses only the content of the analyzed sequences. The main advantage of an intrinsic approach is the ability to find genes not characterized before and not present in transcriptomic analysis. However, gene predictors that use only the intrinsic approach face difficulties to correctly locate the boundaries between introns and exons. Predictor's accuracy can be improved by mixing both approaches, for example, EST information can be used to locate at least some of the exon/intron boundaries, intra-genomic comparisons can help to find multigenic families and inter-genomic comparison can be used to identify orthologous genes. Recently, a new probabilistic model called Conditional Random Field was introduced and it has been used in many areas. We propose the development of a new implementation of a class of Conditional Random Field called Semi-Markov Conditional Random Field that is an extension of our probabilistic framework called ToPS (Toolkit of Probabilistic Models of Sequences). ToPS is a modular computational framework which helps researchers to model, to combine and to experiment probabilistic models. The main advantage of Semi-Markov Conditional Random Field is that it can be used to add extrinsic information to GHMM-like gene predictors. We expect that this new environment will help us develop new, better, gene predictors.

In silico investigation of the contribution of intergenic variations for a behavioral trait in Guzerá cattle

Fernanda Caroline dos Santos, Pablo Augusto de Souza Fonseca, Maria de Fátima Ávila Pires, Izinara da Cruz Rosse, Frank Angelo Tomita Bruneli, Ricardo Vieira Ventura, Maria Gabriela Campolina Diniz Peixoto, Maria Raquel Santos Carvalho

Universidade Federal de Minas Gerais, EMBRAPA Gado de Leite, University of Guelph

Abstract

Aggressive behavior is an undesirable characteristic in dairy herds. Animals characterized as having a bad temperament usually difficult milking procedure, stress other animals, may cause injuries to calves or to farm workers and show worse production quality or rate when compared to calm animals. Therefore, reducing the profits of the herdsman. Behavior is a complex characteristic composed of both genetic and environmental factors. The genetic causes of aggressive behavior are still not well stablished, except for some QTLs or a few genes identified in studies using taurine breeds. In Brazil, most of the cattle is composed of indicine breeds and its crossbreds, being the Guzerá breed one of the most important. In order to identify the genetic variants that contribute to aggressive behavior in Guzerá dairy cattle, 754 female Guzerá were evaluated through REATEST®, an electronic device that converts oscillation and stomping during the weighting routine into a quantitative measure(reactivity). High values of reactivity indicate discomfort, fear and in most of the cases aggressive tendency while the animal is in contention. Whole genome genotypes of these animals were then obtained using the Illumina Bovine SNP50 array, and a genome wide association study (GWAS) was performed for reactivity using GenABLE in R. One of the associated markers, significant in a 5% FDR threshold, was located in an intergenic region, more than 250kb distant of the nearest gene, EPHA6. Considering that the causal variant may be in linkage disequilibrium with this marker, we investigated the region around the marker (800kb upstream to EPHA6) in order to find evolutionary conserved domains (ECR) and transcription factor binding sites (TFBS) inside it, which could be evidence of the presence of regulatory elements of long distance. The ECR browser was used to find ECR cores in this region among cow, fugu, tetraodon, frog, chicken, opossum, mouse, chimpanzee and human. Finally, rVista2.0 was used to look for TFBS inside the ECRs found. Three ECRs were found in the region searched, one of them being conserved in frog, chicken and in all mammals, except for opossum. All ECRs showed recognition sites of multiple transcription factors, which could be involved in regulation of the expression levels of EPHA6, a gene implicated in axon guidance. These results indicate that EPHA6 is a candidate gene for behavioral traits and that regulatory elements of long distance of EPHA6 may contain specific genomic variations that may contribute to aggressive behavior in Guzerá cattle.

In silico gene prediction based on MYOP system

Renato Cordeiro Ferreira, Alan Mitchell Durham

IME-USP

Abstract

In the last decades, many programs were developed with the aim of characterizing genomic genes. Called gene predictors, they usually carry out their function by applying one of two different approaches: intrinsic / ab initio prediction (based only in genomic sequences) or extrinsic prediction (with extra info produced by alignments). Although almost all ab initio predictors reach high levels of precision (PPV) and sensibility (SN) in the classification of nucleotides (>90%), the results for whole exons and complete genes are much lower ($\sim$70% and $\sim$30%, respectively). These numbers are often caused by the difficulty of defining splice sites, which represent one of the greatest challenges for modern gene predictors. In this study, we aimed to measure the performance of MYOP (Make Your Own Predictor), a system for automatic implementation of ab initio gene predictors based on ToPS framework (Toolkit for Probabilistic Models of Sequences). In order to validate MYOP, we compared it with three other prominent gene predictors currently available: Genscan, SNAP and Augustus. We used 6 benchmarks developed by our research group, each containing 2000 sequences of 6 genomes (A. thaliana, H. sapiens, C. elegans, D. melanogaster, Z. mays and O. Sativa). We trained MYOP and Augustus accordingly to a 5-fold cross-validation, and applied SNAP and Genscan with their pre-trained models. For nucleotides, we could notice a balance between MYOP, Augustus and SNAP, while Genscan presented lower statistics for all organisms but H. sapiens. For whole exons, MYOP and Augustus exceeded SNAP performance, reaching differences of more than 10% in PPV for A. thaliana and D. melanogaster. For complete genes, MYOP got the greatest advantage, with the best results in SN and PPV (at least one standard deviation of difference), losing only in H. sapiens sensibility. Through these results, we can include MYOP among the best predictors currently available, capable of predicting as much or more efficiently than Augustus (awarded in gene prediction contests) and with excellent results compared to specialized predictors as SNAP (focused on plants) and Genscan (originally created for H. sapiens).

Characterization and analysis of Corynebacterium pseudotuberculosis respiratory chain and lactate utilization pathway

Carlos Diniz, Elma Leite, Flávia Rocha, Roselane Gonçalves, Mariana Parise, Douglas Parise, Vasco Azevedo, Sintia Almeida

UFMG

Abstract

Corynebacterium pseudotuberculosis is a facultative intracellular bacteria responsible for causing infectious diseases of chronic nature. The disease presents itself in different ways, according to the infected host. This organism has two biovars called equi and ovis, which are classically defined based on their ability to convert nitrate to nitrite in biochemical tests. This project is set in the context of pangenomics, combined with the use of bioinformatics tools. The main objective of this work is to discover the genetic basis and the specific characteristics of the lactate utilization pathway as carbon and energy source, and of the enzimes involved in respiratory chain of each strain of C. pseudotuberculosis. Information obtained from the genome sequence showed that C. pseudotuberculosis possesses a branched electron transport chain to oxygen reducing equivalents obtained by the oxidation of various substrates are transferred to an intramembrane pool of menaquinone-8 via at least five different dehydrogenases. The dehydrogenases include a non-proton-pumping NADH dehydrogenase, malate: quinone oxidoreductase, succinate dehydrogenase, pyruvate: quinone oxidoreductase and L-lactate dehydrogenase. All these enzymes contain a flavin cofactor and, except succinate dehydrogenase, are single subunit peripheral membrane proteins located inside the cell. Em Corynebacterium glutamicum, electrons are passed from the menoquinol via the cytochrome bc1 complex to the aa3-type cytochrome c oxidase with low oxygen affinity, or to the cytochrome bd-type menaquinol oxidase with high oxygen affinity. In C. pseudotuberculosis, were found homologous enzymes to such oxidases terminals de C. glutamicum, concerning this we can consider the existence of this electron transference mechanism to oxygen. Genome analysis revealed that C. pseudotuberculosis possess a operon orthologous lutABC to Bacillus subtilis and Deinococcus radiodurans and a gene lutP upstream, which are abstent in C. glutamicum. The gene encodes a L-lactate permease, whereas the operon encodes L-lactato desidrogenase enzyme which uses quinones as electron receivers and it's attached to the respiratory chain. Multiple alignment performed among sequences of homologous aminoacids related to catalytic subunit of L-lactate dehydrogenase (LutC) enzime of C. pseudotuberculosis and other eight species. Among such species two Corynebacterium genus presented an important sequence similarity, conserved important residues for its structure and catalytic function. Therefore, a better comprehension of C. pseudotuberculosis respiratory chain components improved the knowledge of pathogenicity and survival mechanisms of this microorganism contributing as foundation to future studies and results validation.

Bacteria temperature lifestyle classification using machine learning

Karla Machado, Thais Ramos, Rodrigo Sarmento, Delano Maia, Vinicius Maracaja-Coutinho, Thais Gaudencio

Universidade Federal da Paraíba

Abstract

Identify the adaptations an organism develops over ambiental thermal changes according to its lifestyle is a field of high importance for biotechnological and synthetic biology industry, specially for the understanding of the adaptations for living in inhospitable conditions. The goal of this work is to understand the importance of each nucleotide and its combination, as well as of each amino acid, for the classification of different bacteria according its adaptation to temperature (psychrophiles, mesophiles, thermophiles and hyperthermophiles). For that, we used five different machine learning approaches (k-nearest neighbors, decision tree, naive bayes, neural network and support vector machine), available on Weka tool. All methods were applied on complete genome sequences from 591 bacteria (13 psychrophiles, 479 mesophiles, 77 thermophiles and 22 hyperthermophiles) available on NCBI. The attributes analyzed were: the counts of nucleotides frequencies in groups of two and groups of three for both complete genomes and genes separately; and the counting of individual amino acids. In a pre-processing step, we normalized the nucleotides and amino acids frequencies according to each studied gene, and tested both normalized and non-normalized attributes vectors. Furthermore, we balanced the datasets for each lifestyle according to that with the smaller number of species. Additionally, two groups were generated based on a combination of (i) psychrophiles/mesophiles and (ii) thermophiles/hyperthermophiles; each one of them presenting information related to 99 instances (number of instances of the minor representative class). The training test was developed using the cross-validation of 10 sets. The best results were obtained using non-normalized data and the unbalanced datasets. However, on the unbalanced datasets we obtained a super-adjustment of the model. Thus, considering the unbalanced datasets with two and four classes, respectively, we found an accuracy of 76.92% and 88.89% on the individual counting of amino acids. As already described, these results suggests that the preferential usage of amino acids are determined by its different temperature lifestyle. These results shows the importance of machine learning approaches in order to predict bacterial temperature lifestyle. Currently, we are working on the development of a heuristics to test different groups of parameters for the execution of the predictive methods, and on the relation of amino acids combinations for each one of the four temperature lifestyles.

Detecting bacterial genomic islands using PPM

Paulo Roberto Branco Lins, Karla Cristina Tabosa Machado, Hugo Neves de Oliveira, Vinicius Maracaja-Coutinho, Leonardo Vidal Batista, Thaís Gaudencio do Rêgo

Departamento de Informática, Universidade Federal da Paraíba

Abstract

The DNA is the result of a combination of the genetic material from its ancestors. However, some regions of a particular genome are acquired by horizontal gene transfer from other organisms, resulting in ”alien” regions named as genomic islands (GIs). The acquisition of these new regions results in adaptations to a variety of environment and conditions. This mechanism is one of the main players responsible to bacterial genome plasticity and evolution. In this context, the development of bioinformatics tools for genomic islands prediction is of high importance for bacteria basic and applied research. Here, we used the data compressor algorithm Prediction by Partial Matching (PPM), an entropy codification technique based on the statistics modelling and context prediction, on the identification of genomic islands in different Bacteria. Currently, PPM is considered one of the most effective generic data compressors. It makes use of a group of a maximum of K precedent symbols, for the estimation of the conditional distribution probability for the next symbol from the message. Assuming that genome sequences presents a unique signature, a divergence on the symbols prediction based on the whole genome context may suggest a genomic island insertion. The method was applied on the genome sequence from four different bacteria (Rickettsia prowazekii str. Breinl, Vibrio cholarae chromosome 2 N16961, Vibrio vulnificus CMCP6 chromosome I, Vibrio vulnificus YJ016 Chromosome I), that were previously studied by other groups using different methods for genomic island prediction. It predicted correctly the same island predicted by other methods, and identified other potential new GIs. The already known genomic island located on the position 0,302-0,436Mb from Vibrio cholerae chromosome II was detected by our method, and a potential new island on the position 0.2Mb was found. The same was observed with the other genomes, with all known island identified and at least one potential new island predicted. Finally, our negative control Rickettsia prowazekii str. Breinl - a bacteria known by the absence of genomic islands -, as expected the method did not identified any GI. These results suggests that PPM might be an efficient method for genomic islands prediction. Its efficacy was comproved through the same presence/absence prediction tests available on literature. Additional tests are necessary using different organisms, entropy parameters and other filters for noise reduction.

Polymorphic Endogenous Retroviruses in Primates

Andrei Rozanski, Fabio Navarro, Ana Paula Urlass, Paola A Carpinetti, Anamaria A Camargo, Pedro A F Galante

MOCHSL, Yale University

Abstract

Nearly 8% of the human genome is composed by endogenous retroviruses (ERVs) sequences. It is accepted that ERVs are the result of ancient retroviral infections and that they play a major role in shapping our genome. During ERVs life cycle, homologous recombination takes place. As result the excision of internal region (viral genes: GAG, POL, ENV) and formation of solitary LTRs occur. These viral elements have been related to the development of several diseases such multiple sclerosis, cancer among others. On the other hand, these transposable elements are also related to the gain of new physiologic capabilities. To better understand the role of ERVs, we analysed its evolutive and polymorphic characteristics in primates. We included in our analysis the genome from Human and other five primates (marmoset, rhesus, orangutan, gorilla and chimpanzee). We have selected 5782 LTRs from raw output data obtained from RepeatMasker that is available in GoldenPath. A pipeline to identify orthologous events were developed and applied with success in order to find human specific events. We have detected 200 human specific events, which results in the increment of $\sim$150 events more than previously described. From these events, we developed a customized approach to search for polymorphic events considering 5 superpopulations from 1000 Genomes Project. After that, we were able to characterize 15 polymorphic events LTRs. These polymorphic events showed different allele frequencies among 1000 Genomes individuals. We believe that these results may help to understand the role of ERVs in the structural variation of the human genome. FAPESP support - Project 2013/10659-0

Sequence analysis VH antibodies in mammals: integrating genomic and transcriptomic data

Taciana Conceição Manso, Tiago Antônio de Oliveira Mendes, Liza Figueiredo Felicori

Universidade Federal de Minas Gerais

Abstract

Immunoglobulin heavy chain (IgH) genes are assembled by somatic recombination of VH, DH and JH segments resulting in a population of diverse ligand binding sites in antibodies that protect the organism against infections. The number of V genes encoding potential immunoglobulin parts is highly variable across the different species. The evolutionary diversification and selection processes between different species were poor studied and may help to understand the minimal functional structure as well as key patterns in human antibody essential to design synthetic antibody for diagnosis and therapeutic use. Currently, a large number of genome sequence of different mammals are available allowing exhaustive comparative genomic analysis that will help to understand evolution of IgH gene processing and structure. The goal of this project is to describe the organizational and phylogenetic relationships of sequences derived from the V exons for nine mammals species including humans, mouse, horse, dog, rabbit, sheep, pig and bovine. The genome data of species were obtained in Ensembl database and compared to IMGT/GENE-DB. In parallel, cDNA sequences of the variable region heavy chain available in the NCBI. The genomic motifs and signatures from V genes were recovered and compared to structure of cDNAs from mammals stimulated by different antigens. The sequences obtained were translated in amino acids followed by multiple alignments and the identity percentage for each amino acid from the sequence was calculated based on consensus sequence. Global alignment of transcripts from the same stimulus showed large conserved regions and a few number of variation points generally associated to CDR's. Mapping of each cDNA in conserved and variable IgH regions in the genome was also carried out. Phylogenomics and data mining approach will be applied to these data and we expect to define an evolutionary history of CDR regions and a minimal specific characteristic in human antibodies.

TaxOnTree: a web tool that adds taxonomic classification on top of a phylogenetic tree

Tetsu Sakamoto, José Miguel Ortega

Universidade Federal de Minas Gerais

Abstract

Phylogenetic methods are widely used approaches for analyzing and illustrating protein evolution and are being benefited by the increasing number of species with their genome sequenced. Generating phylogenetic tree with sequences from hundreds of species can be considered as a common task, but, in contrast, tree visualization has been challenged to organize and bring by an accessible means the taxonomic information about the sampled proteins. Here we present TaxOnTree, a web tool that generates phylogenetic tree and focuses on bringing taxonomic information embedded in the tree. TaxOnTree takes as input a single protein identifier and retrieves its putative homologues by BLAST search against RefSeq or Uniprot protein databases. Users can also opt in inputting a list of protein identifiers and proceed to the next steps that are: sequence alignment (MUSCLE or PRANK), inspection of alignment quality (TrimAl) and phylogenetic tree reconstruction (FastTree). Then, taxonomic classification of each species comprising the tree is retrieved from NCBI Taxonomy and taxonomic distances between pair of species are calculated by determining the Lowest Common Ancestor (LCA) between them. All taxonomic information is attached to the tree as a tag in each node generating, in the end, a phylogenetic tree in Nexus format structured to use the tree coloring tools from FigTree software. From the tree file generated by TaxOnTree, besides displaying the basic information about the samples like species name or accession number, taxonomic distances (or LCA) between the query and each subject organisms are easily depicted from the leaf label and from the coloring style of the tree branches. Moreover, TaxOnTree can color the tree branches according to a taxonomic rank (e.g. family, order, class, etc.) selected by the user. This allows users to rapidly inspect the diversity of taxa comprising the tree for a certain taxonomic rank. TaxOnTree provides prompt inspection of taxonomic distribution of orthologues and paralogues. It can be used for manual curation of taxonomic/phylogenetic scenario and coupled to any tool that links homologous sequences to a seed sequence. Thus, TaxOnTree provides computational support to help users inspecting phylogenetic trees with a taxonomic view, even without being taxonomy experts. TaxOnTree is available at biodados.icb.ufmg.br/taxontree. Supported by: FAPEMIG and CAPES

A genetic variability analysis of Basidiomycota ITS, ITS1, and ITS2 regions

Francislon S. de Oliveira, Fernanda Badotti, Aline Bruna M. Vaz, Laila A. Nahum, Guilherme Oliveira, Aristóteles Góes-Neto

Centro de Pesquisas René-Rachou (CpqRR), Fundação Oswaldo Cruz (FIOCRUZ-MG), Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG), Universidade Estadual de Feira de Santana (UEFS), Instituto Tecnológico Vale

Abstract

DNA barcoding is a DNA-based system used to identify previously described species and to facilitate the recognition of new ones, following the general principles of standardization, minimalism, scalability, and rapidity. It ideally utilizes only one standardized DNA segment, which in fungi is the ribosomal internal transcribed spacer (nrITS) region. In specimens with high DNA degradation, such as old herbarium samples, small portions of the barcode region, the mini-barcodes, may be used to substitute the full-length barcode. The same rational can be used when one is intended to carry out a targeted metagenomics analysis of environmental samples. In this case, mini-barcodes are appropriately designated as metabarcodes, since the sample does not comprise a single specimen but a mixture of specimens from distinct species. Basidiomycota is the second most speciose fungal group, exhibiting a striking morphological complexity with a diversity of ecological roles and biotechnological applications. They are important wood decomposers and may be used as food source, in medicine, and in lignocellulose conversion besides many other industrial applications. The fast and reliable identification of these microorganisms is fundamental to many research areas. The main goal of our work was to test the hypothesis of whether the ITS subregions (ITS1 and ITS2) may work as DNA mini/metabarcodes to discriminate basidiomycotan species. In order to reach this goal we have constructed a primary database consisting of all completed vouchered ITS sequences of Basidiomycota currently available in INSCD enriched with metadata, when available, from the UNITE database - `https://unite.ut.ee/`. Quality filters were applied to the primary dataset and then two secondary databases were constructed, one containing ITS1 and the other consisting of ITS2 subregion sequences. Our primary database comprised 8,458 sequences, which represented three subphyla, 25 orders, 76 families, 221 genera, and 1020 species. Detailed and comparative analysis of intra and interspecific variability at the genus level has been performed and will be discussed.

Convergent evolution in enzymes related to antibiotics resistance

Melise Silveira, Antonio de Miranda

Oswaldo Cruz Institute- Fiocruz

Abstract

Convergent evolution can be observed in the anatomy and physiology of living beings, but it occurs mainly at the molecular level. Non-homologous Isofunctional Enzymes (NISE) have no detectable sequence similarity but catalyze the same biological reaction, receiving the same Enzyme Comission Number (EC). NISE are found in all major bacterial lineages, and their number is correlated with the genome size in this domain. The high diversity between bacterial antimicrobial resistance determinants strongly suggests their evolution to the present biochemical function from preexisting molecules with different functions, probably not related with resistance. An example of NISE are beta-lactamases (EC 3.5.2.6); they can use either a catalytic triad or a metal dependent mechanism to hydrolyze beta-lactamics antibiotics, inactivating them. The Ambler structural classification for beta-lactamases is the most used, but it does not represent the evolutionary relationships between these enzymes, as already demonstrated in previous studies. An ideal classification scheme should be objective, stable and predictive, reflecting the current biological scenario, but this does not occurs for the present beta-lactamase classification systems and even does not exist for most antibiotic resistance enzymes. In this work, we have studied a group of 94 curated beta-lactamase sequences to create and implement a pipeline for clustering these sequences, aiming to develop a classification scheme based on sequence similarity, which will reflect their evolutionary relationships. For this, we downloaded the amino-acid sequences from BRENDA, all of them with a reference article describing and characterizing the enzyme. Sequences were clustered using two different approaches: local BLASTclust (with different similarity and coverage values) and an in house script based on best reciprocal hits. After generating the groups, we identified the species and Ambler class for every sequence from every group. We used CD-HIT to eliminate redundant sequences, and then the PROSITE and PFAM databases to identify the motifs and domains present in each group, respectively. Hmmer was used to generate a HMM model for each group. An extensive literature search revealed five databases (CARD, MBLED, Resfams, LacED and Dlact) from where 8990 protein sequences of beta-lactamases could be retrieved. Currently we are running the constructed HMM models against this dataset to check their efficiency in retrieving the appropriate beta-lactamase genes. After this stage is completed, the models will be refined for another interation against RefSeq.

PHYLOGENY AND TAXONOMY MULTILOCUS SEQUENCE ANALYSIS (MLSA) OF THE LEPTOSPIRA GENUS

Elma Leite, César Júnior, Izabela Ibraim, Carlos Diniz, Vasco Azevedo, Flora Fernandes

Universidade Federal de Minas Gerais, Universidade Federal de Pernambuco, Universidade Federal da Bahia

Abstract

Leptospirosis is caused by a bacteria of the genus Leptospira. Currently, Leptospira genus is divided into 20 genomic species, comprehending pathogenic, saprophytic and intermediate species. Based on Leptospira complete genomes and evolutionary aspects of the genus, this study highlights Leptospira species identification by 16S rRNA sequencing. This gene stands as the most used phylogenetic marker for estimating evolutionary history of microorganisms, however, recent studies demonstrate that the 16S gene may eventually be affected by genetic recombination and horizontal transfer. Consequently, phylogenetic analyzes based exclusively on this marker could not reflect the evolution of the genome. Aiming to minimize this effect, taxonomic and phylogenetic studies using MLSA (Multilocus Sequence Analysis), methodology based on combined analysis of multiples genes, has been proposed to bacteria. This work suggests new molecular markers to be used on multilocus sequence analysis of the Leptospira genus. Identification of markers (capD, argK, cheD, metH, LIC20035 e LIC20140) was based on comparative genomics of the following chromosomes II of the following species: L. interrogans serovar Copenhageni str. Fiocruz L1-130, L. borgpetersenii serovar Hardjo-bovis str. JB197 e L. biflexa serovar Patoc strain 'Patoc 1 - Ames, using MAUVE software and BLAST aligner. The concatenated sequences of possible markers were used to construct a MLSA tree. The phylogenetic tree was reconstructed using the PhyMl 3.1 software, while the taxonomic analysis was done using the MEGA6 software. The results obtained utilizing the genetic distance as a criteria for the taxonomic tree were similar to those observed when the maximum likelihood criteria was applied. The MLSA taxonomic tree analysis method demonstrated a monophyletism of the Leptospira group and allowed the phylogenetic position clarification of the strains used in this project. As far as phylogenetic structure, the topology confirms L. biflexa phylogenetic position as the lineage root of the other taxa belonging to the saprophytic bacterium. On the pathogenic bacteria group, it demonstrates L. interrogans as a sister species of L. kirschneri, and L. borgpetersenii as a sister species of Leptospira sp. (Leptospira sp. serovar Kenya str. Sh9). It was also possible to confirm the phylogenetic position of L. kmetyi as belonging to the pathogenic bacterial group. This technical approach has improved significantly the understanding of the genus, as well as showing accurate alternative tools to study taxonomy and phylogeny of prokaryotes.

POTENCIAL MOLECULAR MARKERS FOR TAXONOMY AND PHYLOGENY OF THE LEPTOSPIRA GENUS

Elma Leite, César Júnior, Mariana Parise, Doglas Parise, Vasco Azevedo, Flora Fernandes

Universidade Federal de Minas Gerais, Universidade Federal de Pernambuco, Universidade Federal da Bahia

Abstract

Leptospirosis is a widely distributed zoonosis, caused by pathogenic spirochetes that belong to Leptospira genus. The number of sequenced Leptospira spp. genomes has been increasing continually in the public database, allowing deeper insight into the bacteria biology and evolution. The phylogenetic studies of 16S rRNA gene highlighted that Leptospira genus is divided in two groups, one consisting of saprophytic bacteria, and the other consisting of pathogenic and intermediate bacteria. This study suggests novel alternative molecular markers to be used in taxonomic and phylogenetic studies of the Leptospira genus. To identify these makers, comparative analysis were performed, which can be comprehended in three steps; identification, filtration and validation. In order to identify potential markers MAUVE software was used, these markers were filtered by BLASTn searches and each resulting marker were checked through BLASTp comparisons. These steps were performed to analyze the chromosome II of the following species: L. interrogans serovar Copenhageni str. Fiocruz L1-130, L. borgpetersenii serovar Hardjo-bovis str. JB197 e L. biflexa serovar Patoc strain 'Patoc 1 – Ames. The chromosome II of three species were collectively aligned providing LCBs (Locally Colinear Blocks) and respectively ORFs (Open Reading Frames) visualization. This analysis showed 13 potential molecular makers (capD, LIC20140, LIC20254, pykF, ahcY, acnA, wzb, argK, mcm2, aroK, cheD, metH e LIC20035) of the genus. Each marker was aligned at amino acids level using the MUSCLE software and editing was carried out using MEGA6 software. The best evolutionary model selection for each individual dataset was performed by ProtTest program and phylogenetic reconstruction under the Maximum Likelihood criteria with 1000 bootstrap using PhyML 3.1. The taxonomic analysis distance trees were build using MEGA6. The results showed that the topologies inferred under the Maximum Likelihood and genetic distance criteria associated with multiple external groups suggests a monophyletic genus. In the majority of phylogenetic trees, saprophytic bacterium L. biflexa behave as sister species of L. yanagawae. Concerning to phylogenetic position of the pathogenic group, usually, L. interrogans behave as sister species of L. kirschneri and L. borgpetersenii behave as sister species of Leptospira sp. (Leptospira sp. serovar Kenya str. Sh9). This analysis allowed the selection of six potential molecular markers (capD, argK, cheD, metH, LIC20035 e LIC20140) to be utilized in phylogenetic and taxonomic analysis of the genus.

Molecular identification of green microalgae isolated from Brazilian inland waters reveals putative new species

Sámed Hadi, Hugo Santana, Patrícia Brunale, Taísa Gomes, Márcia Oliveira, Alexandre Matthiensen, Marcos Oliveira, Flávia Silva, Bruno Brasil

Universidade Federal do Tocantins, Universidade Federal de Minas Gerais, Universidade Federal da Bahia, Embrapa Agroenergy, Universidade de Brasília, Embrapa Pantanal, Embrapa Swine and Poultry, Embrapa Amazônia Oriental

Abstract

Traditionally, green microalgae (Chlorophyta) are used as sources of food supplements, pigments and animal feed. There has also been an increasing interest in exploring these microrganisms as sustainable feedstocks for fuels and chemicals. Regardless of that, chlorophyte species identification remains a challenge and usually requires not only morphologic inspection but also physiological as well as DNA-based analysis. Molecular methods for species identification, such as DNA barcoding, provide rapidly and consistent tools for biodiversity monitoring and identification. Currently, however, there is no consensus about the molecular markers (barcodes) that should be used for chlorophytes. The present work evaluated the feasibility of using the Ribulose Bisphosphate Carboxylase Large subunit gene (rbcL) and the Internal Transcribed Spacer 2 of the nuclear rDNA (nuITS2) markers for the identification of a very diverse though poorly known group, the green microalgae from neotropical inland waters. Fifty-one freshwater green microalgae strains isolated from Brazil, the largest biodiversity reservoir in the neotropics, were submitted to DNA barcoding. The sampling areas included natural water bodies within the Amazon rainforest, the Pantanal wetlands and the Cerrado savanna, as well as anthropogenic wastewater deposits. Currently available universal primers for nuITS2 amplification were sufficient to successfully amplify and sequence 90,20% of the samples. On the other hand, novel sets of primers had to be designed for rbcL, which allowed the sequencing of 96,08% of the samples. Thirty-three percent of the strains could be unambiguously identified to the species level based on nuITS2 sequences similarity combined with barcode gap and compensatory base changes (CBCs) calculations. Phylogenetic and morphological analysis confirmed nuITS2-based identification accuracy, including two Desmodesmus and one Micractinium putative new species. In contrast, none of the strains could be reliable assigned to a species based solely on rbcL sequences, since the unavailability of reference barcodes for green microalgae species impaired distance thresholds calculations. In conclusion, the data presented here indicates that nuITS2 should be used as a primary marker, while rbcL should be sequenced as an auxiliary marker to facilitate DNA barcoding of freshwater green microalgae.

The Fate of Duplicated Genes of Cobalamin-Independent Methionine Synthase in Wild and Domesticated Soybeans (Glycine max L.)

Hugo Vianna Silva Rody, Luiz Orlando de Oliveira

Universidade Federal de Viçosa

Abstract

The domestication process can direct strong artificial selection of genes controlling agronomic traits of crops such as soybean (Glycine max), domesticated in China about 4,500 years ago and widely consumed around the world. It has been reported that these genes may exhibit a greater loss of genetic diversity than expected from bottleneck effects as a signature of this type of selection. However, the gene copies (paralogs) of a polyploid organism may encounter different evolutionary fates, depending on which evolutionary forces acted on these paralogs, such as: artificial selection, environmental factors or metabolism processes. We used single nucleotide polymorphisms and existent Next-Generation Sequencing (NGS) data from 18 wild and 14 cultivated soybeans to study how the domestication process affects the genetic diversity and the fate of paralogs by using cobalamin-independent methionine synthase (MetE) as a model gene, because this gene is essential to plant development and nutrition. Statistical tests were applied to compare the pattern of polymorphism, in order to investigate the null hypothesis of neutrality, and to evaluate the occurrence of positive selection in all six MetE paralogs from wild and cultivated soybeans. Our results present preliminary evidence of early stages of neofunctionalization in MetE paralogs and indicate relaxed purifying selection in the other paralogs, due to gene duplication. The greatest genetic variability was between the groups of soybean MetE paralogs and not between wild and domesticated soybeans. Forty sites were predicted to be under positive selection in this protein, indicating that most sites of this protein are under structural constraints in both wild and cultivated soybeans.

Phylogenetic Analysis of Pr1 Proteases in Metarhizium anisopliae

Fábio Carrer Andreis, Augusto Schrank, Claudia Elizabeth Thompson

Universidade Federal do Rio Grande do Sul (UFRGS)

Abstract

Entomopathogenic fungi such as Metarhizium anisopliae (Ma) infect their hosts through cuticle penetration. Transposing this first layer of host defense requires secreted proteases, lipases, and chitinases to degrade its major components, while acting in differentiation of cellular structures and types as well. Our work focuses on the Pr1 family of proteases, which comprises 11 isoforms (Pr1A - K) of two classes: the bacterial-type class I (Pr1C), and the proteinase K-like class II, which is further split in three subfamilies due to shared conserved amino acids and exon/intron composition. Subfamily 1 (Sf1) contains the extracellular Pr1A, B, G, I, and K subtilisins; Sf2 is composed of Pr1D, E, F, and J, also extracellular; Sf3 comprises Pr1H, the only endocellular subtilisin found in Ma. In spite of Pr1 isozyme diversity, they all perform complementary functions in a synergistic fashion for an efficient cuticle proteolysis. In order to assess underlying patterns of evolutionary selection regarding the Pr1 family, we first constructed a local database of 35,560 amino acid sequences based on the Subtilase_S8 PFAM profile using HMMER. We then searched for homologous sequences to each of the Pr1 proteins of Ma strain E6 on our database using BLAST+, and filtered them by identity and size. The resulting sequences were grouped, aligned and manually curated at the subfamily and isoform levels using PRANK (amino acid datasets), and TranslatorX (nucleotide sequences). Afterwards, evolutionary model analyses were performed using ProtTest and JModelTest. Subsequently, phylogenetic analyses were performed using Maximum Likelihood (PhyML), Bayesian Inference (MrBayes), and Quartet Puzzling (TreePuzzle) methods. Each set of trees was compared at the isoform level. The best trees and alignments were tested for positive selection using PAML. Thus far, we constructed phylogenetic trees displaying well-supported clades for each class II subfamily and for most isozymes at the subfamily level. Regarding the Metarhizium genus, branching patterns for most individual isoforms matched that of the currently accepted species tree, which position specialist fungi as primitive to generalists. Since no other genera were observed in this cluster, horizontal gene transfer may not have occurred for Metarhizium Pr1s. We have also identified positive selection in Pr1A, B, I, and K, consistent with the role of these enzymes in host-pathogen interactions. The remaining selection analyses are currently underway. The role of positively selected residues has yet to be determined through structure modeling and molecular dynamics. This work was funded by CAPES, CNPq and FAPERGS.

Phylogenomic study of the segmentation process in flatworm species

Gabriela Prado Paludo, Claudia Elizabeth Thompson, Henrique Bunselmeyer Ferreira

Universidade Federal do Rio Grande do Sul

Abstract

The phylum Platyhelminthes includes all flatworms and comprehends four classes: Turbellaria, Monogenea, Trematoda, and Cestoda. Among flatworms, monogeneans, trematodes and cestodes are exclusively parasites, while most turbelarians are free-living organisms. Some interesting aspects are evident in the evolution of parasitic plathyhelminths, as recent studies have correlated some of their adaptations to parasitic life styles, such as morphological regression and metabolic simplification, with genome reduction in trematodes and cestodes. Moreover, the body formation process (strobilation) is more complex in cestodes than in trematodes. Among cestodes, strobilation apparently evolved in a stepwise pattern, in which serial repetition of reproductive organs (proglottization) and external segmentation (subdivision of proglottides) were independent evolutionary events.The main objectives of this work are the study of evolutionary relationships among segmented and non-segmented species from the Phylum Platyhelminthes, and the identification of genes related to the strobilation process. Considering the sequenced and annotated genomes available in public databanks, 10 parasitic platyhelminth species were included in this study, 5 segmented and 5 non-segmented. A phylogenomic analysis was performed in order to establish their evolutionary relationships, also including genomes from 6 nematodes (non-segmented helminths), one annelid (segmented deuterostome), and one mollusk (non-segmented deuterostome) as outgroups. Ortholog and paralog proteins were identified amongst the deduced complete proteomes, and a set 285 proteins were selected, corresponding to all identified ortholog genes with a single copy in all studied genomes. A supermatrix was generated with the selected protein sequences, and the MrBayes software was used to infer phylogenetic relationships. In parallel, we compared the 10 selected Platyhelminth genomes in order to find 903 exclusive genes in segmented species without orthologs in non-segmented species. In this gene set we have identified 55 development-related based on functional enrichment of proteins using Blast2GO.We will next compare the available Platyhelminth genomes in order to identify genes possibly related to the strobilation process, based on its presence in segmented species and its absence in non-segmented species. With that, we expect to identify at least some genes that are exclusive of genomes of species presenting proglottization associated to external segmentation. Moreover, the available transcription data for fully segmented flatworms will be used to select segmentation-related genes that are differentially expressed in the segmented stages of the life cycles of these parasites. The evolution of these putative segmentation-related genes will be then investigated based on the analysis of positive selection acting on their orthologs in lophotrocozoans.

Phylogenetic analysis of the suckermouth armored catfishes (Siluriformes: Loricariidae) based on mitochondrial transcripts

Daniel Moreira, Maithê Magalhães, Paula Andrade, Paulo Buckup, Carolina Furtado, Adalberto Val, Renata Schama, Thiago Parente

FIOCRUZ

Abstract

The Neotropical family Loricariidae is the most diverse family of catfishes and fifth most species-rich vertebrate family on Earth, containing over 800 valid species. This number is growing fast due to discovery of new species and resolution of cryptic taxa using molecular approaches. This abundance of loricariid species underlies the high diversity of their ecological habitats and high rate of endemism. Despite their extreme diversity and ecological relevance, the evolutionary history of the Loricariidae remains controversial. This study aims to clarify the phylogeny of Loricariidae. The mitochondrial genomes (mitogenome) were assembled from liver de novo transcriptomic data sequenced using the Illumina-HiSeq2500 technology and annotated using bioinformatics tools. This approach is most advantageous because it enables the assembly of almost complete mitogenomes, simultaneously with the sequencing of thousands of nuclear genes. In total, 34 species have their liver transcriptome sequenced and assembled. From those transcriptomes, all mitogenomes were assembled and analyzed from five of six different subfamilies of Loricariidae and one outgroup (Callichthydae). Using this approach, we obtained mitogenomes with coverage ranging from 93% to 100%, comprising the two ribosomal RNAs and the 13 protein-coding genes in all fish sequenced. Only a few of the 22 transfer RNA genes and parts of the control region were missing in some individuals. The gene composition and order were similar to the usual vertebrate pattern. Maximum likelihood phylogenetic analyses using concatenated nucleotide sequences of the rRNAs and 13 protein-coding genes supported Loricariidae as a monophyletic group. However, important differences in the arrangement and composition of the subfamilies have appeared. Our results found support for Hypoptopomatinae including Neoplecostominae, as a monophyletic clade. And this analysis supports a different placement for Loricariinae. Changing places with Hypoptopomatinae clade, which appears as a sister group to Hypostominae. The present study has newly established the almost complete sequence of mitogenomes from 34 species of Loricariidae. Also, our transcriptomic data will be useful to identify nuclear orthologs between loricariids and refine the phylogeny within this highly diversified freshwater fish family. Acknowledgements: Financial support from USAID (PGA-2000003446).

FUNCTIONAL GENOMICS AND EVOLUTION OF ANALOGOUS ENZYMES IN THE HUMAN GENOME

Rafael Piergiorge, Marcos Catanho, Ana Carolina Guimarães

FIOCRUZ

Abstract

Since enzymes catalyze almost all chemical reactions that occur in living organisms, it is extremely important that genes encoding such activities are properly identified and functionally characterized. Several studies suggest that the fraction of enzymatic activities in which multiple events of independent origin have taken place during evolution is substantial. However, this topic is still poorly explored, and a comprehensive investigation of the occurrence, distribution and implications of these events, involving organisms whose genomes have been completely sequenced, has not been done so far. Fundamental questions, such as how analogous enzymes originate, why so many events of independent origin have apparently occurred during evolution, and what are the reasons for the coexistence in the same organism of distinct enzymatic forms, remain unanswered. In this context, the purpose of this project is to investigate the biological importance and the evolutionary role of functional analogous enzymes identified in metabolic pathways annotated in the human genome. A computational pipeline developed by our group (AnEnPi) was used to predict putative analogous enzymes employing protein sequences available in public databases (KEGG and Swiss-Prot). The predicted functional analogy instances will be confirmed by mining in Pfam, SCOP and PDB databases for domain, folding and 3D structure information concerning the enzymes implicated. Using KEGG, Metacyc, RECON X and ENCODE databases as references, the predicted analogous enzymes will be mapped in human metabolism. Finally, a comparative analysis of the expression of genes encoding the predicted analogous enzymes forms will be performed based on RNA-Seq data available in the SRA repository. Thus in this project, we intend to achieve an overview of the mechanisms involved in the de novo origin (convergence) of enzymatic forms in the human species occurred during evolution, as well as to acquire a functional profile of the genes encoding these analogous enzymes, analyzing the transcriptional activities of these genes in different conditions.

Shifts of floral colors in carnivorous plants Utricularia (Lentibulariaceae): a saga told by a phylogenetic perspective

Cristine G. Menezes, Saura R. da Silva, Jackson A. M. Souza, Janete A. Desidério, Rogério F. Carvalho, Vitor F. O. de Miranda

Faculdade de Ciências Agrárias e Veterinárias, UNESP - Univ Estadual Paulista, Câmpus Jaboticabal, Instituto de Biociências, UNESP - Univ. Estadual Paulista, Câmpus Botucatu

Abstract

Anthocyanins, a class of flavonoids, are the main flower pigments in plants and play important roles in attracting insects for pollination and also protecting the organs from photodamage. The flavonoid biosynthetic pathway is an interesting example of a pathway that has evolved gradually expanding as new products with new functions were added. Considering the relation of floral color and pollinator attraction, any modification in the flower may result in the pollinator shift or negative selection of floral phenotype. The color modifications in the flowers are confidently traced in phylogenetic hypotheses and occur in reason of mutation of genes that encode important enzymes of the flavonoids pathway or also by the gene suppression caused by transcription factors. Here we present phylogenetic history of color flowers in the carnivorous plant Utricularia (Lentibulariaceae) based on the plastidial DNA sequences: trnL-F, rbcL and rps16. Carnivorous plants from the genus Utricularia are distributed world-wide and comprise approximately 235 species occurring across every continent except Antarctica. They are highly specialized plants with very specialized and modified leaves adapted to capture and to digest prey – usually small arthropods. Seventy-nine species (some species with more than one individual, for instance U. amethystina that was represented by the three phenotypes: purple, white and yellow-flowered plants) of Utricularia were analyzed, based on sequences from NCBI and also sequenced to this study. Genlisea species were used as out-group. Maximum parsimony trees were produced with PAUP version 4b10 (heuristic searches with 2,000 replicates). The maximum likelihood analyses were produced with RAxML BlackBox software with the clades supported by bootstrap. To the Bayesian inferences we used the MrBayes version 3.1.2. The best-of-fit model was calculated with jModeltest version 2.1.1 and based on corrected Akaike information criterion (AICc). To Utricularia the plesiomorphic color is purple, and the floral colors red, yellow and white are apomorphic states and derived from the purple color. The floral color is highly homoplastic considering all species of the genus, with some bias: the shifts from purple to yellow or to white are up to 6 times more frequent than the reverse (white to purple or yellow to purple). Our data support the hypothesis that the genes from the base of the anthocyanin pathway (e.g. CHS, CHI, FLS or DRF) were modified by deleterious mutation or maybe by transposons, thus the yellow and white (not pigmented) flowers could be a result of partial or total suppression of these genes.

Phylogenomic investigations of plant pathogenicity in the fungus responsible by witches' broom disease on cocoa trees

Juliana José, Gustavo Costa, Paulo Teixeira, Daniela Thomazella, Gonçalo Pereira, Marcelo Carazzolle

Unicamp, University of North Carolina, UC Berkley

Abstract

Phylogenomic approaches for the investigation of genomes reveals a key information beyond comparisons by similarity: the evolutionary process underlines the genomic patterns we observe. In a biological system usually investigated by its impact on cocoa production, we propose an evolutionary perspective on the genomic analysis of the plant pathogen fungi Moniliophthora perniciosa, responsible for the witches' broom disease on Theobroma cacao. In the present work we conducted a phylogenetic-based comparison of M. perniciosa with 12 species related to the Agaricales order level at least, using public available genomes. In order to understand what kind of genomic patterns may be related to the development of pathogenicity pathways specific to host cocoa trees, we have intentionally chosen species with different life-styles: biotrophic, saprotrophic and endophytic. The phylogenetic relationships among Moniliophthora and other Agaricales species were established using orthologs among all species, through bayesian methods implemented on BEAST, providing strong branch support (> 0.9) for the whole topology. Using the phylogeny and the homology assignment of gene families among all Agaricales obtained with OrthoMCL, the amount of gene gain and losses in each ancestral lineage were estimated using the maximum likelihood method implemented in CAFE and evidences of positive natural selection were investigated by likelihood estimates of dN/dS in codeml. Moniliophthora presented quite fewer exclusive gene families than other Agaricales indicating that its species retain more ancestor gene characteristics. However, the Moniliophthora lineages still have shown a larger proportion of gene gain than gene losses within its families. Pathogenicity-related families with significant gene gain also presented evidence of directional natural selection as the main processes driving their evolution. Other gene families with significant gene gain will be further studied.

Origin of genes obtained by transcriptomic data compared to KEGG Functional Hierarchies

Katia Lopes, Ricardo Vialle, J Miguel Ortega

UFMG

Abstract

To study the origin of genes we analyzed data of the Human Genome Consortium Fantom version 5 and functional families of KEGG BRITE hierarchies. Firstly, the FPKM data from Fantom5 were converted to TPM data. Next, we determined the occurrence of homologues of human genes by inspecting the database KEGG Orthology (KO), enriched by us with UniRef50 clusters. Thus we set up to determine the time of appearance of subset of genes along human evolution. There are a total of 56 tissues in our local database. They present expression levels for 21,097 genes mapped to 16,567 human proteins. Accessing the group of orthologs of these genes and determining the LCA (Lowest Common Ancestor) of them, we determined the fraction shared with all nodes of human lineage. Our data show that human share 14% of their genes with bacteria and archaea and 47% with eukaryotes. In the same period, the protein Kinases of KEGG shared only 0.87% with bacteria and 30

Ligand-Based Pharmacophore Modeling and Virtual Screening of Ligands for the Lanosterol 14-Alpha Demethylase Protein from Leishmania infantum

Natalia Fonseca, Nilson Nicolau-Junior

Universidade Federal de Uberlândia

Abstract

Leishmania infatum is one of the species that causes visceral leishmaniasis, one of the most serious forms of this disease that could be lethal if not treated properly. The treatment, currently available for visceral leishmaniasis, is associated with many side effects, complex therapeutic regimen, high cost, and, sometimes, it do not result in cure. This study aims to search for new compounds with pharmacological potential against Leishmaniasis, by the search for ligands to the three-dimensional structure of the enzyme lanosterol 14-alpha demethylase (CYP51) from Leishmania infantum (PDB ID: 3L4D). Computational tools were used to construct a pharmacophore model based on the fluconazole, a known ligand of the CYP51 protein. The model was used in subsequent steps of validation and virtual screening using ligand libraries. The pharmacophore modeling was performed with the aid of ROCS 3.2.0.4 (OpenEye Scientific Software, Santa Fe, NM). This model contains information about shape and chemical properties extracted from the fluconazole molecule. The ligand libraries used in this research are originated from two Chembridge Corporation datasets, specifically, the DIVERSet and EXP, that have been carefully selected, totaling 100,000 drug like compounds. In order to perform the virtual screening, the ligand libraries were prepared with the OMEGA 2.5.1.4 (OpenEye Scientific Software, Santa Fe, NM), which was used to generate conformer libraries. The pharmacophore model was previously validated using the ROC (receiver operating characteristic) curve and AUC (area under the curve). The ROC curve presents the comparison between bioactive compounds and decoys (ligands unrelated to the target protein and potentially inactive). The AUC extract from the ROC curve graph it is simply the probability that randomly chosen bioactive compounds have a score higher than randomly chosen inactive compounds. In order to generate the ROC curve and the AUC value, biologically active ligands against CYP51 were obtained from ZINC database and the decoys were generated on the DUD-E online platform. The ROC curve graph generated had an AUC value of 0.8, the minimum expected for a good pharmacophore model. This result showed that the model had high sensitivity or, in other words, it presented a preference to the biologically active ligands higher than the decoys. After validation, the conformer libraries previously generated were submitted to the pharmacophore model and the top 500 ligands of each, based on the TanimotoCombo score, were selected. The best-scored ligands will be used to perform a molecular docking with CYP51 from Leishmania infatum.

Computational study of statin derivative with biological activity against HMG-CoA reductase (HMGR) using Molecular Docking.

Jéssica de Oliveira Araújo, Rutelene Natanaele Barbosa de Sousa, Heinrich dos Santos Menezes, Clauber Henrique Souza da Costa, Amanda Ruslana Santana Oliveira, João Elias Vidueira Ferreira, Ricardo Morais de Miranda, Antonio Florêncio de Figueiredo

Instituto Federal de Educação, Ciência e Tecnologia, Universidade Federal do Pará, Universidade Federal do Pará

Abstract

Hypercholesterolemia (> 240 md/dL) is a disease characterized by high pathological level of low density lipoproteins (LDL), which take cholesterol from liver to bloodstream. High LDL levels may form atheroma, which may cause cerebral vascular accident (CVA), aneurysm and heart attack. The disease affects 30% of people in Brazil, particularly those older than 45 years old and is the main cause of death in the United States of America. One of the synthetic drugs used to treat hypercholesterolemia are statins, which inhibit enzyme 3-hydroxy-3methyl-glutaryl-CoA reductase (HMGR). The action of this enzyme occurs particularly in the liver by inhibiting the conversion of the substrate into mevalonic acid that is the precursor of cholesterol. Because of the similarity between mevalonite and B-lactonic structures of statins, these compounds bind the active site of the enzyme HMGR to compete with HMG-CoA. This way cholesterol production is suppressed. Compactin, simvastatin, fluvastatin, cerivastatin, atorvastatin and rosuvastatin are the most important statins that are capable of keeping acceptable cholesterol levels in the blood. The most common side effect associated to statins is myalgia. It is believed that statins with higher hepatoselectivity can help to reduce this side effect, because the compounds are less available to the muscle tissues. This work describes a theoretical study of a compound from the family of statin that presents biological activity against enzyme HMGR. The molecular structure had its geometry optimized using B3LYP method with 6-31+G* basis set implemented in Gaussian 03 software. Molecular electrostatic potential (MEP) maps were generated through Molekel 5.4 software in order to help to find key features related to the biological activity investigated. Furthermore molecular docking was performed with aid of AutoDock Tools 4.2 program using the chains A and B of the complex 1HW9 extracted from the Protein Data Bank (PDB). After molecular docking the structures were compared with the crystallographic ligand of simvastatin to analyze the similarity and the shape of the statin compounds bind. This structure showed an excellent molecular docking, which suggests the free energy of binding is favorable to inhibit the enzyme HMGR considering the statin derivative studied.

Modeling and Molecular Dynamics of the largest subunit of the Ribulose-1,5-Bisphosfate Carboxylase/Oxygenase (RuBisCO) from Cyanobacterium Limnothrix sp. CACIAM 53.

James Siqueira Pereira, Andrei Santos Siqueira, Leonardo Teixeira Dall'Agnol, Juliana Simão Nina de Azevedo, Evonnildo Costa Gonçalves

Universidade Federal Rural da Amazônia - UFRA, Universidade Federal do Pará - UFPA

Abstract

Ribulose -1,5- bisphosphate carboxylase/oxygenase (EC 4.1.1.39 , RuBisCO) is the most abundant protein in the world and is considered the major enzyme involved in the photosynthesis process. RuBisCO is classified in four distinct forms: Forms I, II, III and IV. RuBisCO form I is a hexadecameric protein structure with eight copies of both large and small polypeptides in an (L2)4(S4)2 structure codified by rbcL e rbcS genes, respectively. This form is the predominant RuBisCO found in nature and it is present in Cyanobacteria, algae and plants. To unravel the structure and function of this enzyme in Cyanobacteria, this study aimed to construct a three-dimensional model (3D) of the large subunit of a Cyanobacterium from the Amazonic Collection of Cyanobacteria and Microalgae – LTB/UFPA. Amino acid sequence was obtained from a genomic study of Limnothrix sp. CACIAM 53 isolated from superficial water of Tucuruí Hydropower Plant Reservoir, Pará State, Brazil. The mold selection was chosen using Blast tool included in the PDB database, bringing the best identity with RuBisCO from the Synechococcus PCC6301 (PDB ID: 1RBL.A). The three dimensional structure was generated through Modeller 9.10 and subsequently validated by the Ramachandran plot, Verify3D, Anolea and the Root Mean Square Deviation (RMSD). Additionally, the evaluation of surface's electrostatic potential was performed through a map construction generated by PBEQ solver server. Finally, Molegro Virtual Docking was used for an analysis of molecular dynamics (MD) to evaluate the substrate in the catalytic site fitting. The obtained structure showed 15 β-sheets and 19 α-helix. The Ramachandran plot showed 97.4% of residues within energetically favorable regions and 93% of residues showed positive value in the 3D-1D evaluation. Individual residues analysis done by Anolea resulted in a few regions with high energy and the obtained RMSD value was 0.173. The map of electrostatic potential revealed similarity between the molecules regarding to their charge distributions, even to the active site region with low electron density. The best conformation obtained in MD process showed the main interactions already described, highlighting those with Lys167, Lys169 and His290 residues, as well as with magnesium ion. The highest structural conservation, including electrostatic charges and interactions presented by the obtained model, classifies it positively, been contributing to studies that aims to optimize the carboxylase activity of RuBisCO and cyanobacteria biomass exploitation.

IN SILICO ANALYSIS OF KEY ACTIVE SITE RESIDUES AND CHARACTERIZATION OF THE HIUASE/TRANSTHYRETIN PROTEIN FAMILY BY DECOMPOSITION OF RESIDUE CORRELATION NETWORKS

Natan Pedersolli, Lucas Bleicher

Federal University of Minas Gerais

Abstract

The Transthyretin/HIUase protein family constitutes a remarkably case of divergent evolution, from an enzyme to a hormone transporter. HIUases are found in all kingdoms and are part of the purine pathway, catalyzing the conversion of 5-hydroxyisourate (HIU), the final product of uricase, into 2-oxo-4-carboxy-5-ureidoimidazoline (OHCU). Transthyretin is a plasma protein found only in vertebrates, secreted by the liver and choroid plexus, which transports thyroid hormones (T3 and T4) and retinol. Furthermore, mutations in TTR may result in protein misfolding, which can lead to amyloid cardiomyopathy and neurodegenerative diseases. More recently, it was revealed that transthyretin and HIUase presents possibly biologically relevant zinc binding sites, a view that was reinforced by the detection of metal-dependent peptidase activity in the transthyretin family. In this work we analyze the current set of known sequences for the Transthyretin/HIUase using decomposition of residue correlation networks and structural analysis of the key residues for enzymatic activity, hormone binding and protein-metal interactions. Due to the lack of a crystallography data of HIUase in complex with 5-hydroxyisourate, we use receptor-ligand docking methods to analyze the behavior of key residues obtained previously in complex with the ligand and intermediary compounds of the HIU hydrolysis. Our analysis shows that aside for the expected correlations, there is also a strong signal mapping to a charged-pair interaction between residues 63 and 87 (E. coli numbering) and the importance of arginine and histidine for the 5-hydroxyisourate binding and enzymatic activity. We conclude that although the highly biased distribution of proteins in the sample (the vast majority of homologs are HIUases, due to its older origin and highest distribution across the kingdoms, while transthyretins are much less abundant and also more similar), it is still possible to obtain useful data from conservation and correlation analysis, if proper measures are taken.

Computational study of kojic acid by inhibiting glyoxalase I enzyme against leishmaniasis using molecular docking

Rutelene Natanaele Barbosa de Sousa, Jéssica de Oliveira Araújo, Heinrich dos Santos Menezes, Clauber Henrique Souza da Costa, Amanda Ruslana Santana Oliveira, João Elias Vidueira Ferreira, Antonio Florêncio de Figueiredo, Ricardo Morais de Miranda

Instituto Federal de Educação, Ciência e Tecnologia do Pará, Universidade Federal do Pará

Abstract

Leishmaniosis is a very serious infectiuous disease but that is neglected. The disease is caused by the genus Leishmania sp and is transmitted by phlebotomus (mosquitos that feed on blood) from the genus Lutzomya. The disease may be either a zoonosis or an anthropozoonosis or both. There are two types of Leshmania: cutaneous and visceral and symptoms are cutaneous. According to WHO, life conditions are associated to the disease. People living in poor sanitary conditions and in areas next to forests and even with frequent climate changes arein risk of contracting the disease. The protozoan of lytic cycle attacks directly the cells that protect the host (macrophage), reproducing itself inside the cells and causing damage to it and, consequently, infection in the host. The protozoan has in its cytoplasm an enzyme called glyoxalase I (a metalloprotein from the glyoxalase complex, essential to the protozoan's survival) responsable for the detoxification of the toxin methylglyoxalase that is released to face the protozoan. Literature points that many inhibitors (good flavonoids) may act against glyoxalase enzyme. In this work kojic acid (KOJ), which is already used in dermatology, is investigated in silico against glyoxalase I enzyme. First the molecular structure of KOJ had its geometry optimized using B3LYP method with 6-31+G basis set implemented in Gaussian 03 software. Then molecular electrostatic potential (MEP) maps were generated through Molekel 5.4 software in order to help to find key features related to the biological activity investigated. Furthermore molecular docking and was performed with aid of AutoDock Tools 4.2 program and calculations based on Fukui index to better understand the interaction between KOJ and the active site of the protein.

Cruzain pharmacophore modeling and virtual screening

Viviane Correa Santos, Rafaela Salgado Ferreira

UFMG

Abstract

Cruzain is the major Trypanosoma cruzi cystein-protease and is related to parasite nutrition and host cell invasion. Its inhibition is shown to decrease parasite infection in animal models but no medicine has been developed yet. Cruzain is an human cathepsins homologous, and we have to be aware of it when developing drugs against this enzyme. In order to develop a selective inhibitor to cruzain we propose to model pharmacophores with the software LigandScout in a structure-based mode. We will model cruzain, cathepsin L and B pharmacophores and apply a parallel virtual screening with the ZINC database drug-like compounds. First, we will screen the compounds against a pharmacophore query and then we will dock the hits against the enzymes for a more accurate pose prediction. We searched in PDB crystal structures of cruzain, cathepsin L and B co-crystallized with inhibitors to model the pharmacophores. We found 24 cruzain hits, 28 to cathepsin L and 9 to cathepsin B. We clustered cruzain ligands to remove redundancy and found 3 clusters and 3 singletons. Selection of the cluster representative ligand was done by analyzing the electron density map of the PDB files with the Coot software. To validate the pharmacophore model we need some active ligands and decoys. We obtained 65 actives through literature search and 3772 decoys in DUDE database. Several pharmacophore models were already obtained, but they are being refined by evaluating the area under (AUC) the Receiver operating Characteristic (ROC) curve to each alteration we do. We compare the AUC values to each model and the higher will be selected to be the screen query. The perspectives includes docking the pharmacophore hit compounds against the enzymes of interest.

A machine learning approach to detect enzymes participating in lipid metabolism pathways of bioenergy plants based on protein sequence properties

Rodrigo Oliveira Almeida, Ney Lemke, Guilherme Targino Valente

UNESP - Botucatu

Abstract

The worldwide growing concern about environmental issues is inducing changes in fossil fuels usage policies. An interesting alternative are biofuels. Renewable energy source, biodiesel can be obtained from different oil crops. However, to a large use of this biofuel, is necessary to develop high lipid concentrations sources. In the last years, biologic data generated are increasing quickly and current is necessary tools more effective to analyze this high amount of data. Thereby, machine learning is a interesting tool to help analysis of lipid metabolism data. The present study aim to look for patterns on enzymes related to lipid metabolism in four bioenergy plants (Glycine max, Sesamum indicum, Arachis hypogaea e Brassica napus) in attempt to construct models that could be useful for other species. The data set (protein sequence) was obtained from ocsESTdb (`http://ocri-genomics.org/ocsESTdb/`) e Uniprot (`http://www.uniprot.org/`) and using programming language R to structure the data set. Duplicated sequences was removed and two groups was generated, the enzymes of lipid metabolism (E-L) and no enzymes of lipid metabolism (E-NL), which they have 3,202 and 35,204 protein sequences, respectively. Using the Peptides and ProtR packages, 1,402 attributes was generated to both groups. It was created 11 training data sets (6,404 instances for each one) after undersampling the E-NL group and one test data set (38,406 instances). All data set was submitted to Weka software, using the J48 algorithm and 10-fold cross validation. The mean of correctly classified instances and AUC was 78.13% and 80.05%, respectively. To obtain a representative model, was performed a hierarchy clustering of the 11 models generated, which allowed to select just 33 attributes present from 6 models clustered. These new data sets were submitted to Weka (J48) and resulting in a mean of 78.09% and 79.59% of correctly classified instances and AUC, respectively. The results show that these 33 attributes were enough to perform the data classification, specially the attributes length of protein and percentage of basic amino acids of the protein. Furthermore, the final results not were harmed. Supported by: CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior)

A simple procedure to determine ligand specificity – application to Ring hydroxylating oxygenases

Lucas Carrijo, Lucas Bleicher

Universidade Federal de Minas Gerais

Abstract

Polycyclic aromatic hydrocarbons (PAH) are organic compounds formed by two or more fused aromatic rings. Due to their carcinogenic and mutagenic potential, they are considered important environmental pollutants. One technique commonly applied in order to minimize the impacts caused by PAH is bioremediation. It consists in using microorganisms that possess enzymes capable of metabolizing these compounds. Ring hydroxylating oxygenases (RHO) are enzymatic complexes responsible for the first step of the degradation pathway. It consists in a reductase, a ferredoxin and an oxygenase itself. The last one, in turn, is composed by three alpha and three beta chains. Alpha chain shows two distinguishable domains: an iron-containing catalytic domain and a Rieske domain containing an iron-sulfur complex. Although non-genetically modified organisms are capable of accomplishing this task, some efforts can be made to enhance the efficiency of these enzymes. For instance, one can look at which amino acid residues confer specificity for a given substrate to the enzyme. Likewise, one can look up the reasons why an enzyme is more promiscuous than another. Therefore, the aim of this work is, once we know these features, try to modify the enzymes in such a planned way that could improve their activity. For this purpose, the full alignment of Ring Hydroxyl A catalytic domains (PF00848) was downloaded from the PFAM database and filtered in order to remove fragments, poorly aligned sequences and redundancy. Then, a set of UniProtKB/Swiss-Prot entries of alpha chains from enzymes with different specificities was used as references. For each Swiss-Prot sequence in the Pfam alignment, their residues were sorted by decreasing order of conservation in order to build a curve showing the size of a sub-alignment containing N of the most conserved residues of the reference sequence. In general the curve looks like a decreasing sigmoidal function, in which the first plateau represent the residues that are conserved in the full alignment, and the second one represent those residues that are exclusive for the reference sequence. Using this procedure, those specificity determining residues can be extracted from the curve slope. By comparing curves of sequences with the same specificity, it is possible to infer which of those less conserved positions are good candidates to be specificity determinants. The most remarkable observation in this analysis is that only about ten residues are function diagnostics, while the majority of residues are sequence-specific.

Amino acid correlations in the Low Molecular Weight Phosphatases protein family

Marcelo Querino Lima Afonso, Lucas Bleicher, Priscila Graziela Alves Martins

Universidade Federal de Minas Gerais

Abstract

Protein Tyrosine Phosphatases (PTPs) are enzymes responsible for regulating the removal of phosphate groups from tyrosine residues in proteins. Class II phosphatases, composed by Low Molecular weight phosphatases (LMW-PTPs), exhibit a unique folding pattern and have no sequence similarity to other phosphatase classes beside the P-loop, a CX5R motif responsible for phosphate ion binding. Structurally related to these enzymes are a group of Arsenate Reductases, which catalyse Arsenate reduction in Bacteria by a reaction mechanism that involves two other cysteine residues besides the highly conserved catalytic cysteine in the P-loop. In our work, we applied our group's framework for finding amino acid correlations in the PFAM alignment consisting of LMW-PTP fold proteins (PF01451). The results yielded seven residue communities and a large number of anti-correlations. We then analysed all the family published structures and revised the literature in order to describe possible community functions for all the correlated residue pairs. In one of these communities, we observed correlations between residues important for substrate binding and catalysis and P-loop conformation maintenance during the reaction. Two other communities corresponded to Arsenate Reductase catalytic cysteine residues and another community possessed a tyrosine residue targeted for phosphorylation and involved in LMW-PTPs Protein-Protein interactions. A community of particular interest consists of a P-loop cysteine important for redox regulation in multiple phosphatase classes by forming a disulphide bridge with the catalytic P-loop cysteine and a P-loop glycine whose importance is previously undescribed in the literature besides substrate hydrogen bonding. We now plan to analyse the influence of this glycine residue on this disulphide bridge formation by analysing P-loop conformations using Steered Molecular Dynamics.

Revealing protein-ligand interaction patterns through frequent subgraph mining

Alexandre Victor Fassio, Sabrina Azevedo Silveira, Carlos Henrique da Silveira, Raquel Cardoso de Melo-Minardi

Universidade Federal de Minas Gerais, Universidade Federal de Viçosa, Universidade Federal de Itajubá

Abstract

Molecular recognition plays an important role in biological systems and is a phenomenon of organization very difficult to predict or design even for small molecules. Due to its remarkable importance, molecular recognition was studied under different perspectives in Bioinformatics. Understanding and predicting protein-ligand interactions, although these are complex tasks, are essential steps towards ligand prediction, target identification, lead discovery and drug design. In this work, we are interested in receptor (protein) and ligand (non-protein) interactions which consists of non-covalent bonding such as aromatic stackings, hydrogen bonds, hydrophobic interactions and salt bridges. Therefore, we propose a model and algorithms to understand why different small molecules are recognized by a specific protein. Our model is based on graphs. Each protein-ligand complex is a bigraph where nodes are atoms and edges depicts interactions between protein and ligand. The proposed algorithms are based on graph mining and aim to search and detect conserved subgraphs in the dataset of graphs representing protein-ligand complex interactions. In this approach, conserved subgraphs would represent emerging patterns responsible for protein-ligand interaction and molecular recognition. However, the mining process can generate a exponential number of patterns so that the analytical process will be extremely toilsome as an expert has to assay each of the patterns carefully and they can be very voluminous. Accordingly, we also propose a visual interface where users can find general statistics about the dataset, the type of atoms and interactions established as well as select and analyze the generated patterns. The use of images to represent information is becoming more and more appreciated for the benefits it can bring to science by providing a powerful means both to make sense of data and to communicate. We show an example of use of this methodology with Ricin and CDK datasets, both with their respective ligands. In both instances we were able to confirm experimental results from literature.

A sequence-structure atlas for Class I Protein Tyrosine Phosphatases

Mélcar Collodetti, Priscila Graziela Alves Martins, Néli Fonseca, Lucas Bleicher

UFMG

Abstract

Most signaling pathways in eukaryotes involve the reversible phosphorylation of proteins, a process which is controlled by the concerted action of protein kinases, which add a phosphoryl moiety to a protein residue (in most cases tyrosine, serine or threonine) and phosphatases, which removes such group. While protein tyrosine kinases (PTKs) form a single homologous family, protein tyrosine phosphatases (PTPs) are a complex set of families from different evolutionary origins. Particularly, Class I phosphatases are composed of two distinct, but homologous, groups – the Classical PTPs, which dephosphorylate tyrosine residues only, and Dual Specificity PTPs (DSPs), which can also dephosphorylate serine and threonine. Due to their importance to disease related processes, PTPs are largely studied by structural biology techniques: there are currently about 330 available three dimensional structures from classical PTPs and180 from DSPs on the protein data bank. In order to better obtain specific information from the available biological data, we present a large scale analysis on Class I PTPs based on multiple sequence alignment statistics and structural analysis. Using conservation analysis and decomposition of residue coevolution networks (DRCN), we identify residues which are strictly necessary for the structure and function of Class I PTPs and also those which are less conserved but act collectively for specific characteristics.These positions are then interpreted in the light of known literature about protein tyrosine phosphatases – mostly structural and mutagenesis-based studies. In order to facilitate the visual interpretation of our results to a broader public, we also provide a set of modified PDB files and scripts for Pymol which allows a user to interactively analyze relevant PTPs and generate graphical representations of overall conservation and coevolved residue sets.

THEORETICAL STUDY BY MOLECULAR DOCKING OF POTENTS ANALOGUES OF METHOXYLBENZOYL-aryl-THIAZOLES IN THE FIGHT AGAINST OVARIAN CANCER

Renan Patrick da Penha Valente, Clauber Henrique Souza da Costa, Amanda Ruslana Santana Oliveira, Luiz Eduardo Valente Monteiro, Antônio Florêncio de Figueiredo, João Elias Vidueira Ferreira, Ricardo Morais de Miranda

Instituto Federal de Educação, Ciência e Tecnologia do Pará, Universidade Federal do Pará

Abstract

Cancer is the name given to more than 100 diseases that have in common the chaotic growth of cells that, by quickly division, can be uncontrollable and causes serious damages, such as the formation of malignant tumors and also spreading in other regions of the body. When expressed in the ovary, cancer becomes a difficult gynecological tumor to be diagnosed, so it is unlikely to cure, because it is usually already discovered in an advanced stage. The latest global estimative indicates that there were 238.000 new cases of ovarian cancer in 2012, with an estimated risk of 6.100 cases per 100.000 women. The highest incidence rates can be observed in western and northern parts of Europe and North America. However, there are ways to combat it, and one of the treatments is through drugs that act on these cells, more specifically in the cytoskeleton, on the tubulin protein, forming the microtubules, that are fundamental in the process of cell division. There are two models of action of these chemotherapy drugs; one of these is by the stabilization of microtubule polymerization and another is by destabilization, called antimitotic. This review focuses on the analysis of analogues of Methoxylbenzoyl-aryl-thiazoles, a compound derived from colchicine, with resistance factors much more satisfactory than drugs currently used in chemotherapy of ovarian cancer, which act on the destabilization of the microtubule, interacting in the same binding site, stopping the multiplication of cancer cells and ceasing its existence. First, there was a quantum mechanical treatment of molecular optimization with B3LYP/6-31G** method and calculations of molecular docking with the macromolecule crystallographic 1SA0 (Colchicine-Tubulin Complex) removing the PDB (Protein Data Bank), made with AutoDockTools 4.2 program thus analyzed the properties such as distance of interaction between ligand-receptor and binding free energy, in order to describe the behavior of the analogues at the active site of the protein. The docking results showed that the derived substances form binds between hydrogen and cysteine residue (Cys241B) of tubulin protein, having a favorable binding energy, and conformational structure of molecules was similar to colchicine structure found in the active site of the protein.

FUNCTIONAL ANALYSIS OF THE CONFORMATIONAL EFFECTS OF MUTUALLY EXCLUSIVE ALTERNATIVE SPLICING EVENTS

Julio Nunes, Andrea Balan, Tiago Sobreira, Paulo Oliveira

Universidade de São Paulo, Purdue University, Centro Nacional de Pesquisa em Energia e Materiais

Abstract

The mutually exclusive alternative splicing (MEAS) has a unique and significant contribution to the structural diversity of proteins, which results in a wide possibility of affinities and interactions among the isoforms. These changes not only enable new conformations with other important and critical molecules, but also favor the increase of new pharmacological strategies. However, the functional diversity of these protein isoforms is due to the presence of interactions between their modified domains. Given the above, it is important to determine the functional effects of such events. In this sense we use a comparative methodology applied to the secondary, tertiary and/or quaternary conformations of these protein isoforms. To that end, we performed a comparative analysis of the main molecular features inherent to the different conformations of the pairs of protein isoforms originated by MEASs. Also, we identified the mechanisms involved in such events thereby elucidating the target process and the effects that maintain these pairs as molecular modulators. Throughout this process we systematically compared and analyzed the domains, binding sites and functional residues located in regions affected by MEAS. Finally, to support other experimental evidence of the involvement of MEAS with alleged etiopathogenesis, we used various databases of functional data. After a curated analysis of the final set of representative templates, we obtained a sample of 101 events present in 64 genes. Out of this number, we emphasized the high incidence of kinases involved in the regulation of functionally diverse etiopathogenesis, among which we highlight the cancers. Among other findings, the structural results suggest that the MEASs: (i) modulate the homodimerization of certain kinases, (ii) modify the selective interactions with their protein substrates thereby mediating apoptosis and proliferation, (iii) regulate broad domains, (iv) modify the conformations which would be used by the pockets of functionally important molecular substrates, (v) regulate constitutive domains functionally important for ubiquitination and protein degradation, and (vi) generate a regulation of phosphorylation. These groups of gene isoforms analyzed encode the proteins seriously involved in the chemotherapeutic treatment of various cancer diseases, especially those modified by selective interactions with their substrates. Taken together, our findings support our emphasis on presenting a group of isoforms modulated by MEAS and amenable to therapeutic use through feasible molecular regulatory innovations. This work has financial support by CAPES.

Flexibility study the Dengue protease using normal modes and molecular dinamics

Patricia Cassiolato Tufanetto, Antônio Sérgio Kimus Braz, Luis Paulo Barbour Scott

UFABC

Abstract

Dengue virus is responsible for approximately 300 million infections each year worldwide. It is transmitted through an arthropod vector from the genus Aedes, being Aedes Aegypt the most commom. The clinical infection can develop in the classical manner or more serious conditions - dengue hemorrhagic fever (DHF) and dengue shock syndrome (DSS) - that are associated with hemorrhages and volume depletion and can be fatal. It is by far the most devastating of all the recognized arthropod-transmitted virus diseases. However, despite numerous efforts, currently there are no specific approved measures for the treatment or prevention of dengue - the only measure so far is to control de arthropod vectors. In this sense, there is a crucial interest in the characterization of drug targets against DENV. The DENV genome encodes a polyprotein precursor which is processed by a protease, giving rise to three structural proteins and seven nonstructural (NS). The proteases is a serine proteases with a trypsin-like fold that contains two domains: NS3 and NS2B. It is believed that the protease domain, located at the N terminal region of the NS3 interacts with domains of the NS2B protein that contribute to the substrate recognition region of the protease. However there are different conformations of the NS2B cofactor available in previous X-ray and NMR structures and it is not possible to comprehend the catalytical competence and biological significance of the folds without further studies. This project should provide information about motions of the viral protease whose main features are a large amplitude and low frequency . This type of motions is usually associated with mechanisms of allosteric regulation. In addition, this information will be very useful in future studies of protein x Docking protein and virtual screening for the development of new active site and allosteric inhibitors.

In-silico analyses for the discovery of drug and vaccine targets in Burkholderia cepacia: A Novel Hierarchical Approach

Sandeep Tiwari, Syed Babar Jamal, Syed Shah Hassan, Debmalya Barh, Artur Silva, Vasco AC Azevedo

Universidade Federal de Minas Gerais, Institute of Integrative Omics and Applied Biotechnology, Federal University of Para

Abstract

Burkholderia cepacia complex (BCC) is a pathogen usually causing infection to immuno-compromised or hospitalized patients. It is also associated with infections in patients with underlying lung disease, such as cystic fibrosis and chronic granulomatous disease. In 2003, Welcome Trust Sanger Institute sequenced the first genome of Burkholderia cepacia complex strain (i.e. B. cenocepacia J 2315). Pathogen genome sequencing and comparative genomics have resulted in identification of large number of effector genes shown to be responsible for promoting pathogenesis in plant cells. The effector genes clustered in pathogenicity islands (PAIs) can be identified by scanning the genome regions for atypical GC content, codon usage biased approaches and other nucleotides statistical analysis. The present study aims at identification and qualitative characterization of promising drug targets in Burkholderia cepacia using a novel hierarchical in silico approach, encompassing three phases of analyses. In phase I, four sets of proteins were mined through chokepoint, pathway, virulence factors, and resistance genes and protein network analysis. These were filtered in phase II, in order to find out promising drug target candidates through subtractive channel of analysis. The analysis resulted in therapeutic candidates, which are likely to be essential for the survival of the pathogen and non-homologous to host, human anti-targets, and gut flora. Finally, in phase III, the candidate targets were qualitatively characterized through cellular localization, broad spectrum, interactome, functionality, and druggability analysis. The study explained their subcellular location identifying drug/vaccine targets, possibility of being broad spectrum target candidate, functional association with metabolically interacting proteins, cellular function (if hypothetical), and finally, druggable property. Outcome of this study could facilitate the identification of novel antibacterial agents for better treatment of Burkholderia cepacia infections.

In silico analysis drug and vaccine target identification using subtractive genomics against Streptococcus agalactiae, strain GBS85147

Edgar Lacerda de Aguiar, Sandeep Tiwari, Syed babar Jamal, Vasco Ariston Carvalho Azevedo

PG program in Bioinformatics (LGCM), Institute of Biological Sciences

Abstract

Streptococcus agalactiae strain GBS85147, is a gram-positive and bacterial pathogen. This species can causes diseases in humans, cattle and fishes. In humans, it is associated with neonatal sepsis and meningitis, as early-onset or late-onset diseases (EOD, LOD). Although being a common colonizer of the gastrointestinal and genitourinary tracts, it can also affect immunocompromised adults. In dairy cattle, GBS is an important pathogen of clinical and subclinical mastitis, affecting quality and production of milk. In fish, S. agalactiae is an emerging pathogen that causes septicemia and meningoencephalitis with high mortality in wild and cultured species worldwide. S. agalactiae is a bacterium of great medical and veterinary importance due to a high social and economic impact, together with the number of pathogenicity in different hosts. The incidence of invasive infections unrelated to pregnancy in human adults and animals seems to be increasing worldwide, justifying an increase in the number of studies in the area. Currently there is no effective drug or vaccine available against S. agalactiae GBS85147. To identify new targets, we adopted a subtractive genomics strategy, using currently sequenced genome Streptococcus agalactiae strain GBS85147 from our research group, we identified 3 potential targets which were essential and non-host homologs (considering cow, fish and human as hosts) satisfying all criteria of being putative therapeutics targets. Additionally, we subjected these 3 proteins for virtual screening with compound library obtained from natural sources. In all cases, molecules were predicted to form favorable interactions which showed high complementarity to the target, were found among the top ranking compounds.

Sequence and structure-based analysis of Rieske domains from proteins with bioremediation activity potential

Juliana Silva, Lucas Carrijo, Lucas Bleicher

Instituto de Ciências Biológicas - UFMG

Abstract

Population growth is accompanied by increasing demands on oil, which results in a large release of organic pollutants – such as polycyclic aromatic hydrocarbons (PAHs) – in nature. An effort for diminishing such contaminants is bioremediation using enzymes or microorganisms capable of metabolizing them into less aggressive compounds. Pseudomonas putida is a gram-negative bacteria with great metabolical versatility and low nutritional needs, being able to degrade many PAHs. This is achieved by an enzymatic complex composed by a reductase, a ferredoxin and an oxygenase. The oxygenase itself is an heterohexamer consisting of three alpha and three beta subunits. The alpha subunit has two domains – the catalytical domain, which presents an iron ion, and a Rieske domain, which presents an iron-sulphur cluster. In this study, we analyze the different domains of this protein in order to investigate how the reaction occurs in the structure, which residues are related to specificity and how this enzyme could be possibly modified by computational modeling. We used decomposition of residue coevolution networks (DRCN) to detect functionally and structurally important residues in Rieske domains obtained from the PFAM database. The x-ray structure of naphthalene 1,2-dioxygenase from Pseudomonas putida was used for structural analysis, and also for detection of highly frustrated regions using the Frustratometer server and normal mode analysis using Prody. DRCN analysis identified residues which according to its structural localization and prevalence in the protein family may be key to catalysis, and we have detected that the small region in the enzyme containing the iron-sulphur cluster is simultaneously highly frustrated and extremely mobile when the protein chain is analyzed individually, a behavior which is diminished upon the oligomeric complex formation. Given that this region is close to an active site in the chain it binds to, it is likely that the presence of a small set of residues in this region and its characteristic dynamic properties may be as important for catalysis as the active site itself, which lies in the other domain. We plan to further investigate this region by molecular modeling of point mutations, which may alter these characteristics, and comparison of three-dimensional structures from homologous enzymes with different specificities.

Using HMMs and protein motif patterns to generate decoy sequences for bioinformatics teaching

Dhiego Andrade, Lucas Bleicher

UFMG

Abstract

Hidden Markov models (HMMs) have many applications in biology and, more specifically, in biological sequence analysis. They are a very powerful tool to detect distant homology between biological sequences, with profile HMMs being the driving force behind PFAM, the most comprehensive database of protein domains. PFAM catalogs protein domains by starting from seed alignments which are used to build profile HMMs, and these are then used to screen large protein sequence database (such as Uniprot) to detect all putative homologs, which are then made available as multiple sequence alignments. Simpler sequence motifs, such as those associated with post-translational modifications or metal binding, can be represented by much less complex models such as regular expressions or even ordinary consensus sequences. In bioinformatics or biochemistry disciplines, it is common to expect students to be able to extract information from a protein from their sequence alone, which can be done by running them through sequence analysis servers, doing database searches and/or finding or modeling three dimensional structures. However, the large availability of automatically annotated sequences means that the actual process of discovering information from a novel sequence cannot be reproduced if students use existing sequences. Here we present a method to generate decoy sequences which can be used for such assignments, combining the generation of larger protein-like sequences from PFAM domains HMMs and further inclusion of smaller motifs into such sequences. We believe this way the information abstraction may be more effective. This approach will be used in the undergraduate courses for Bioinformatics and Protein Biochemistry at UFMG.

Prediction of druggable proteins based on dipeptide frequency

Marcio Luis Acencio, Gaurav Kandoi, Ney Lemke

Institute of Biosciences of Botucatu, São Paulo State University (UNESP), Iowa State University

Abstract

The emergence of -omics technology has fuelled the progress of rational drug design and drugscreening methods. Developmental costs however increase at a much higher pace than the number of approved drugs. Lack of efficacy, low hit-to-lead ratio and an unclear understanding of disease at the molecular level have forced experimental researchers to incorporate computational methods early in the process. Over the years, there has been a great shift in the paradigm of predicting druggability with the focus of scientific community being shifted towards machine learning approaches. Given a set of labeled dataset, machine learning algorithms like Naive Bayes, Support Vector Machines, Decision Trees and Ensembl methods are able to learn class-specific properties. In the present work, we harness the potential of simple sequence properties, specifically dipeptide composition, to identify potential druggable proteins. We built a decision-tree based meta-classifier by using a training set of 825 known druggable proteins from the Therapeutic Targets Database and calculated their dipeptide composition using an in-house Python script. Our meta-classifier is able to recover 88% of known druggable proteins with a precision of 86%. Moreover, the probability of a protein predicted as druggable belongs to the set of known druggable proteins is 95%. Finally, by training the J48 algorithm with the druggability dataset we generated decision trees to discover rules for druggability. According to these decision trees, the most frequent di-amino acid pairs are formed essentially by hydrophobic amino acids. We believe that this would help in prioritizing potential drug targets and would accelerate the pace of drug discovery.

MOLECULAR DOCKING STUDIES OF PEPTIDE DERIVATIVES INHIBTORS OF CRUZAIN WITH BIOLOGICAL ACTIVITY AGAINST CHAGAS DISEASE.

Adria Perez Bessa Saraiva, Beatriz Silva Quaresma, Biatriz Ferreira de Moraes, Clauber Henrique Souza da Costa, Amanda Ruslana Santana Oliveira, João Elias Vidueira Ferreira, Antonio Florêncio de Figueiredo, Ricardo Morais de Miranda

Instituto Federal de Educação, Ciência e Tecnologia do Pará, Universidade Federal do Pará

Abstract

Chagas disease is an infection caused by the protozoan Trypanossoma cruzi. The vectors that transmit the disease are triatominae, hematophagous insects, popularly known as barbeiro, which are infected by the parasite. Transmission occurs when someone scratches the area bitten by the insect and its excrement, having the trypomastigote form, penetrates the skin and reaches the bloodstream. According to the World Health Organization around 7 to 8 million people were infected by T. cruzi in over 20 countries in Latin Amarica in 20014. It is believed that 8 million people infected with the disease can put into risk 100 million people. Among the drugs investigated, only two active compounds showed favorable action against the disease: nifurtimox and benznidazole. Both are efficient during the acute phase of the disease, but they differ in relation to the efficiency against string. Then it is extremely necessary to develop a new therapy to find more potent drugs. Cruzain is the major cysteine protease of T. cruzi and is essential to the survival of the parasite in cells of the host and consequently to spread the Chagas disease. This enzyme was pointed as a target for potential inhibitors. The major classes of inhibitors of this enzyme include peptide derivative that showed both in vitro and in vivo activity. In this work a computational study was performed on the peptide derivatives with biological activity against T. cruzi. So the peptide derivatives had their molecular structure geometry optimized using B3LYP method with 6-31G* basis set implemented in Gaussian 03 software. After molecular docking with the crystallographic macromolecule 1AIM (cruzain) was performed with aid of AutoDock Tools 4.2 program to better understand the interaction between the inhibtor and the protein. Bond distances were computed between the atoms in the inhibitor and the atoms attached in aminoacids found in the active site in the receptor looking for favorable free binding energy (FBE). Results revealed that favorable FBE through hydrogen bond with GLY66A, suggesting agreement between molecular docking and experimental results. This way inhibitors presented good conformations, something that demonstrates that they are effective against cruzain, opening possibilities to develop new drugs against Chagas disease.

A COMPUTATIONAL STUDY OF PEPTIDIC INHIBITORS OF CRUZAIN.

Beatriz Silva Quaresma, Adria Perez Bessa Saraiva, Biatriz Ferreira de Moraes, Clauber Henrique Souza da Costa, Amanda Ruslana Santana Oliveira, Luiz Eduardo Valente Monteiro, João Elias Vidueira Ferreira, Antonio Florêncio de Figueiredo, Ricardo Morais de Miranda

Instituto Federal de Educação, Ciência e Tecnologia do Pará, Universidade Federal do Pará

Abstract

The Chagas disease, caused by Trypanosoma cruzi protozoan, is presented in two main phases: acute and chronic. The blood-sucking insect of the Triatominae family, better known as Barber is responsible for transmitting the disease to the mammalian hosts and after the blood meal, deposits their feces containing the epimastigotes and trypomastigotes metacyclic forms of the parasite near the hole used for suction. The penetration of the parasite in the form of metacyclic trypomastigotes is possible by scratching or mucosa, resulting in the subsequent invasion of the host cells, such as macrophages, fibroblasts and smooth and striated muscle cells. According to the WHO (World Health Organisation) about 6 to 7 million people are infected worldwide, especially in Latin America, where Chagas disease is endemic. In Brazil, they are recorded between 150 and 200 cases of the disease annually by the Ministry of Health data. In many drugs tested only two compounds were effective in the treatment, nifurtimox and benznidazole, both active only in the acute phase of the disease. The cruzain, a primary T. cruzi cysteine protease is essential for parasite survival in host cells and therefore is an important target for the development of inhibitors as potential therapeutic agents. According Yundchool Choe et al., enzyme inhibitors include peptide derivatives that have shown activity in vitro and in vivo against the parasite. For this reason, it is essential to the discovery of new drugs for the proper treatment of Chagas disease. This work begins with the study of macromolecule crystallographic cruzain (1AIM) extracted from the Protein Data Bank PDB (Protein Data Bank). Initially the compounds were subjected to molecular optimization calculations with B3LYP / 6-31G method, then were carried frontier orbitals HOMO and LUMO calculations and Molecular Electrostatic potential maps (MEP), viewed with the aid Molekel software 5.4, in order to obtain information on the electron-donor and electron-acceptor character of the compounds and electron density of the molecule, used to predict the ability to attract or repel other molecules. Therefore, this study will provide some of the important information you will need to understand about the interactions between the inhibitors and cruzain.

A genetic algorithm to identify functional signature based on physicochemical characteristics

Larissa Leijôto, Raquel Minardi

Federal University of Minas Gerais

Abstract

Comprehension of structure, function and mechanisms of protein are some of the most difficult tasks in Bioinformatics. Therefore, diverse methods have been proposed to determine conserved patterns among proteins expecting that these patterns will help in a protein understanding. Similarities can manifest themselves in diverse levels of proteins and they can be found through comparisons among sequences and structures, highlight conserved and variable regions helping to determine functional and evolutionary relationships. Alignments are the most popular techniques to find conserved stretches in proteins and to spot differences or specificities. However, proteins are extremely complex molecules, which may have distant evolutionary relationship what makes them unique. Those distant evolutionary relationships, usually lead to a weak sequential identity what makes sequence align methods fail to detect relevant patterns. Superposition, or structural alignment methods, in turn could be alternatively more effective because structures are known to be much more conserved than sequences. Nevertheless, superposition methods cannot discover homologous proteins regions that have different foldings. In this way, these types of techniques also are susceptible to fail in specific contexts, which are important. We proposed a multiple alignment method based on varied structural features of proteins in order to detect structural signatures. We are based on the fact that methods that use structural information are recognized in the literature for improvements in the alignment quality, and consequently their accuracy. We aim to avoid weak parts of the alignment sequence and structure superpositions, proposing a methodology based in a genetic algorithm where individuals are representatives of each protein in a database, and the structural features are introduced in representation of each sequence of amino acids. The fitness of an individual in our genetic algorithm is determined by a similarity function among structural features, thus our objective is to find the largest subsequence that has similarity according to structural characteristics. These characteristics are based on the contacts that amino acids establish with their neighbors in tridimensional space, and they are weighted by strength that these contacts exert. We believe that this methodology is more appropriate to embrace characteristics that are really relevant in protein comparisons and on the identification of a functional signature.

Theoretical investigation through Molecular Docking of Triclosan derivatives in the inhibition of the FAS-II trans -2-ACP-enoil reductase (ENR) to fight Malaria.

Heinrich dos Santos Menezes, José Cleyton Nascimento Glins, Jéssica de Oliveira Araújo, Rutelene Natanaele Barbosa de Sousa, Clauber Henrique Souza da Costa, Amanda Ruslana Santana Oliveira, Luiz Eduardo Valente Monteiro, João Elias Vidueira Ferreira, Antonio Florêncio de Figueiredo, Maria Solange Vinagre Corrêa, Ricardo Morais de Miranda

Instituto Federal de Educação, Ciência e Tecnologia do Pará, Universidade Federal do Pará, Universidade Federal do Pará

Abstract

Malaria or paludism is a disease caused by unicelluar parasite Plasmodium falciparum, transmitted by the bite of the female of Anopheles mosquito. There are over 100 species of Plasmodium, however, only four affect humans. After transmission, symptoms such as malaise, fatigue, classical fever and weakness may develop quickly in some cases by up to eight days. According to the World Health Organization (WHO), malaria cases decreased from 227 million in 2000 to 198 million in 2013. The estimated cases per 1000 inhabitants at risk showed a 30% decrease in the global incidence and 34% in the African region of WHO. If the annual rates of reduction in the past 13 years are maintained, malaria mortality rates are projected to decrease by up to 55% globally and 62% in the African region. In the Americas, Brazil is the country with the highest incidence, being prevalent cases in the Amazon region, according to Ministry of Health data, the disease has reached about 265 thousands of people, a decrease of 23% over the previous year. Triclosan, a specific inhibitor of FAS-II pathway trans-2-enoyl-ACP reductase (ENR, also known as InhA or Fabi) is effective against a variety of bacteria, widely used as an antimicrobial agent in formulations for domestic use, including soaps and toothpaste. And recently it was discovered Triclosan's ability as a specific pathway inhibitor FAS-II in P. falciparum to the fight against malaria. Therefore, this study conducted a theoretical research of Triclosan analogs, performing a quantum mechanical treatment of molecular optimization with B3LYP/6-31G and Molecular Docking calculation method in the chain of crystallographic macromolecule 3AM5 (Triclosan complex) the PDB withdrawal (Protein Data Bank), made with the aid of AutoDockTools 4.2 program, and the properties were taken as a distance of interaction between ligand-receptor interaction and free energy to describe the behavior of the analogues at the active site of the protein. The docking results showed a high similarity with the binder crystallographic Triclosan, so these studies are in agreement with experimental results.

Prediction of Secondary Structure of Proteins using Logistic Regression

Carmelina Leite, Lucas Bleicher, Marcos Augusto dos Santos

Federal University of Minas Gerais

Abstract

The prediction of protein secondary structure is a line of research in the field of bioinformatics utilizing large number of methods and have great importance in various fields, for example, therapeutic and biotechnological. Their advance produces directly impacts on health and knowledge in biological processes. Despite the achievements and advances, the prediction of protein structure remains a challenge. By applying the logistic regression statistical technique will be possible to predict the secondary structure of the protein by assigning the triplets residues the probability values of the occurrence of a structure (Beta-sheet or alpha-helix). This new technique can integrate the methods of ab initio protein structure prediction. It is an innovative method that seeks to use a statistical technique already established for predicting protein secondary structure. Currently, there are several techniques that are utilized to ab initio predictions, however, can only be achieved without molds predicts new proteins of about 100 residues. The intention of this study is to predict the most likely structure for several triplets, and consequently, increasing the complexity of the protein that hasn't template. However, this method may not work for specific cases where the influence of the neighborhood has more influence than the sequence. We propose a secondary structure prediction method using regression Logistics. A significant sample of the database Protein Data Bank (PDB) was made and processed the data using the MatLab software version 7.10.0.499 (R2010a). For each fragment with size 9 residues, we use a sliding window of size 3 to characterize each. Each triplet has a probability value associated with the occurrence of secondary structure in question. A model for alpha-helix structure and tape-beta was built. By obtaining the list of triplets assigning probability to with one, it will be possible to predict the secondary structure of the protein in question. At present, the model was only tested for the alpha helix secondary structure. The results are very promising, and can be improved in the future

Molecular modeling triclosan derivatives with biological activity in the combat against Malaria

José Cleyton Nascimento Glins, heinrich dos santos menezes, Jéssica de Oliveira Araújo, Rutelene Natanaele Barbosa de Souza, Luana Priscilla Ribeiro Seki, Clauber Henrique Souza da Costa, Amanda Ruslana Santana Oliveira, Antonio Florêncio de Figueiredo, João Elias Vidueira Ferreira, Maria Solange Vinagre Corrêa, Ricardo Morais de Miranda.

instituto federal de educação ciência e tecnologia do Pará, universidade federal do Pará

Abstract

High fever, sweaty, myalgia, headache and malaise are some of the symptoms of malaria, a parasitic disease caused by protozoa of the genus Plasmodium. The Plasmodium falciparum parasite is transmitted by the bite of the female Anopheles mosquito, in rare cases, the disease can spread through blood contact between infected and uninfected. Some individuals may exhibit seizures and icterus symptoms, which are the rarest type of disease, cerebral malaria, responsible for about 80% of lethal cases. According to the World Health Organization (WHO) in 2013, about 198 million cases were globally reported, it is estimated that 584.000 deaths occurred in Africa, representing 90% of cases of death. In August 2015, WHO has pointed out the high number of deaths caused by malaria, reaching 600,000 cases, emphasizing the urgency with which the authorities should take steps to reduce these numbers. Triclosan proved to be an important pathway inhibitor FAS II, more specifically FAS II trans-2-enoil–ACP reductase (ENR, or also known as InhA) and is highly useful in combating various microorganisms, such as: Escherichia coli, mycobacteria and multiresistant Staphylococcus Aureos. Recently, Triclosan has been tested as an inhibitor of P. falciparum. The study was conducted by administering the compound to lab mices infected with Plasmodium berghei, which has very similar behavior to P. falciparum in humans. The Triclosan subcutaneous administration 3mg / kg ratios for 4 days resulted in a 75% reduction in parasitemia. In mice that had already acquired an infection, a single dose of 38 mg/kg resulting in a total clearance of the parasite, even at higher dosage, liver and kidney functions were normal. The results show that Triclosan and its analogues can be converted into suitable drugs to combat malaria. This computational study begins with molecular optimization calculations with B3LYP method with 6-31G * basis set and Molecular Electrostatic potential map calculations (MEP). MEP maps were visualized with help of the program Molekel 5.4 and assist in the investigation of key features responsible for the biological activity of Triclosan derivatives in the fight against malaria. The information obtained with MEP method about the local polarity of derivatives assist in the study of the regions of interaction of the compounds of 3AM5 protein (Triclosan complex) withdrawn from the PDB (Protein Data Bank), clarifying fine patterns of interaction.

Computational Study and Inference of Mutations Affecting Viral Fitness and Escape from Immune System in the HIV-1 Envelope Glycoprotein

Amanda Albanaz, Carlos Rodrigues, Douglas Pires, David Ascher

CPqRR, Fiocruz - Minas Gerais, Universidade Federal de Minas Gerais/Brazil, University of Cambridge/United Kingdom

Abstract

Mutations play a primordial role in evolution, however changes on protein structure can affect its stability, function and interfere with its interactions with natural ligands or drugs, giving rise to drug resistance or reduce their efficacy. Combating multi-drug resistant strains in virus and bacteria is a major challenge to public health, giving the ever increasing number of cases worldwide. In this scenario, the ability to computationally predict the impact of missense mutations on protein structure and their interactions is strategic for predicting the emergence of drug resistance. The main goal of this work is to propose a new in silico method for predicting resistance mutations in antigen/antibody complexes and their evasion mechanism of the immune system. In particular, we aim to study mutations on the HIV envelope glycoprotein 120 (gp120) in complex with human antibodies. We have collected experimental data for 119 mutations on the structure of gp120 in complex with different human antibodies, regarding neutralization sensitivity and binding affinity. Homology models of gp120 in complex with three human antibodies (VRC01, CD4, b12) were created using MODELLER. The mutations were mapped on these structures and the contacts of the wild-type residues in calculated with the program Arpeggio. Descriptors were calculated for each mutation, including stability and protein-protein affinity change predictions (using the programs DUET and mCSM-PPI, respectively), solvent accessibility and distance to interface. These, combined with the experimental phenotype, were used as evidence for training predictive models using regression trees. While predicting antigen-antibody binding affinity changes of the VRC01-gp120 complex our method achieved a correlation of $R^2 = 0.73$ on training and $R^2 = 0.56$ on leave-one-out cross-validation (LOOCV). This increases to R^2=0.67 after 10% outlier removal on LOOCV. While predicting neutralization sensitivity on the CD4-gp120 complex our method achieved a correlation of $R^2 = 0.71$ on training (and $R^2 = 0.51$ on leave-one-out cross-validation. These, however preliminary, are encouraging results that show how mutation effects can be modeled with a concise set of descriptors. Hereafter we intend to improve the predictive model by correlating other descriptors to phenotype. We believe the proposed approach is flexible and can be applied to different system and disease scenarios where drug resistance is a concern, including tuberculosis and cancer.

A MULTI-DEPENDENT SIDE-CHAIN ROTAMER LIBRARY FOR PROTEIN STRUCTURE PREDICTION

Bruno Borguesan, Marcio Dorn

Federal University of Rio Grande do Sul

Abstract

Rotamer libraries are commonly used to correctly assign the dihedral side-chain angles for amino acid residues in protein structures. Rotamer libraries are widely used to assist the problems of Structural Bioinformatics like the protein structure prediction, protein design, structure refinement, homology modeling, and X-ray and NMR structure validation. These libraries are mostly classified as backbone-dependent, backbone-independent and secondary-structure-dependent. The first group are libraries that consists of rotamer frequencies, mean dihedral angles, and variances as a function of the backbone dihedral angles. The second group makes no reference to backbone conformation, but use side-chain information from all experimentally determined protein structures available. The third group present rotamer frequencies and dihedral angles for each secondary structures, such as alpha-helix, beta-sheet and coil. However, even when these romater libraries are used an enormous possibility of side-chain angles can be allowed. To reduce the complexity of the side-chain search in the Tertiary Protein Structure Prediction problem, we propose a novel approach that combine backbone-dependent, amino-acid-dependent and secondary-structure-dependent. We study the side-chain conformational preferences of 6,650 protein structures. From that point of view, we combined the side-chain angles of each amino acid residue in the secondary structure with the frequencies of backbone dihedral angles. Based on this analysis, we developed a rotamer library that combine all these conformational preferences to assign the dihedral side-chain angles for amino acid residues in protein structures. This multi-dependent rotamer library was combined and tested with a Knowledge-based genetic algorithm for the tertiary protein structure prediction problem. Preliminary results show that the proposed library can help knowledge-based protein structure prediction methods to improve their predictions

Web interface to apply energy minimization of globular proteins in cloud environment

Lucas Exposto, Alexandre Defelicibus, Rodrigo Faccioli

Faculdade de Barretos, Universidade de São Paulo

Abstract

The proteins have an important role, because many of the functions of life are carried out by them and, then, the study of its structures allows elucidate their functions and properties in terms of molecular vision. The complexity and heterogeneity of the data in Computational Biology require, for an effective use of Computational Biology, the use of a computing infrastructure that can handle and manipulate the data in an efficient and effective way. As noted in a report of 2013 sponsored by the US National Science Foundation (NSF) in Arlington, the cloud environment provides access of computational resources for the researchers to carry out the required softwares to obtain results of research. In addition, such softwares are available to attend the demand of both I) end-users that without the technical knowledge about computing, as well as to ii) expert developers who develop of softwares for biological research. More specifically, the process to minimize the energy of the three-dimensional arrangement of globular protein to perform molecular dynamic simulation qualifies with an example of this requirement. Here, it is proposal a web interface in cloud environment to perform energy minimization of globular proteins. This web interface is composed by integration of several softwares whereas allow to perform two kind of energy minimization: I) minimization without restrictions is applied steepest descents algorithm for allowing protein conformation reached a minimum ”naturally” and ii) minimization with restrictions all-bonds is used steepest descents algorithm to preserve all bonds of protein conformation. In this work, we have used the Gromacs, in which it relates to softwares for minimization of protein energy, as well as, Bioinformatics Galaxy that permit to build a web interface to final user and share the results.

In Silico identification of inhibitors against ribose 5-phosphate isomerase from Trypanosoma cruzi

Vanessa Sinatti Luiz Phillippe Baptista, Ernesto Caffarena, Ana Carolina Guimarães

Fundação Oswaldo Cruz

Abstract

Chagas' disease is a public health problem in Latin America, including Brazil. According to the Pan American Health Organization (PAHO), around eight million people in the Americas are infected with Trypanosoma cruzi. Despite this situation, the drugs used for treating this disease are active only in the acute phase, presenting low efficiency and many side effects. Therefore, new-therapeutic targets with critical importance for the parasite must be used in the development of new drugs that are more effective and less toxic to humans. In this context, the enzyme ribose 5-phosphate isomerase (R5PI) is a promising molecular target, since the R5PI of T. cruzi (TcR5PI) and H. sapiens are analogous. The R5PI enzyme catalyzes the reversible reaction between D-ribose-5-phosphate (R5P) and D-ribulose-5-phosphate (5RP). R5PI plays a central role in the pentose phosphate pathway that has been proposed to be crucial in the protection of T. cruzi against oxidative stress, as well as in the production of nucleotide precursors and NADPH. In this work, we have used molecular docking techniques to identify potential inhibitors retrieved from the ZINC database. We analyzed the active site of the X-ray structure of TcR5PI (PDB Code 3K7S) and other similar structures from the Protein Data Bank (similarity larger than 50%) to determine potential structural waters. Subsequently, redocking studies were carried out using AutoDock Vina and GlideXP with the crystal structure of TcR5PI prepared with and without the critical water molecules in the active site. The AutoDock Vina redocking performed employing the crystal structure containing the conserved waters in the active site showed a slightly better match of docked and the crystallographic binding orientations with an RMSD of 1.52 and binding energy of -7.0 Kcal/mol, while the GlideXP redocking showed better results without the conserved waters (RMSD of 0.39 and docking score of -10.240). Virtual screening was performed using the same crystal structures of the redocking step and compounds from ZINC database, which. was screened for structurally similar compounds to R5P (ring and linear forms) and 5RP and by the Lipinski's rule of five. The presence and lack of conserved waters in the active site of the TcR5PI crystallographic receptor were crucial for good redocking outcomes depending of the molecular docking software. The virtual screening based on the best conditions generated different subsets of potential drug candidates. Thus, perspectives are to perform in vitro experiments to validate the top rated compounds in both approaches.

Identification of molecular targets in Trypanosoma cruzi using compounds derived from beta-lapachona: A cheminformatics approach

Luiz Phillippe Baptista, Vanessa Sinatti, Ana Carolina Guimarães

Fundação Oswaldo Cruz

Abstract

The Chagas disease is a public health problem in Latin America, with a strong impact in Brazil. According to the Pan American Health organization (PAHO) about 8 million people in the Americas are infected with Trypanosoma cruzi. Rate of 12 thousand deaths per year. Despite this, the drugs used for the treatment are only active in the acute phase, have low efficiency and many side effects. Thus, the development of more effective drugs that are less toxic to humans is needed. In this context, three promising compounds derived from naphthoquinones (named N1, N2, and N3) were tested by Menna-Barreto and collaborators showing high trypanocidal activity in all forms of T. cruzi and low toxicity to mammalian cells. Previous proteomic analyzes indicated differences in the protein expression profile when the parasites were treated with these three compounds, providing tracks of possible pathways in which these compounds could be acting. However, the targets of these compounds (and their mechanisms of action) are yet unknown. The main objective of this work is to identify the proteins involved in the parasite response to the compounds N1, N2, and N3. These targets could be used in structure-based drug design. In this work, we used cheminformatics methods in order to search for T. cruzi protein targets and where these compounds could be acting, aiming to corroborate the analysis obtained in vitro. We analyzed the N1 compound using SuperPred, PharmMapper, PASS ONLINE, HitPick, Spider, and DRAR-CPI servers. From the wealth of data obtained we could find interesting patterns: according to the server Spider, the N1 compound could interact with DNA Topoisomerase. This indication is very interesting since some naphthoquinones (notably lapachol) are known to inhibit Topoisomerase I and II; PharmMapper suggests Tyrosine-protein kinase as a target for the N1 which also agrees with the literature. Plant-derived naphthoquinones were described as weak inhibitors for c-Src protein tyrosine kinase. Finally, we will perform the same tests with the other two compounds. The results obtained will be analyzed in order to reach a consensus and, hopefully, the final targets will guide a structure-based drug design.

Structural impact analysis of missense SNPs present in the uroguanylin gene

Antonio Marcolino, Állan Pires, Leonardo Ferreira, William Porto, Sérgio Alencar

Universidade Católica de Brasília

Abstract

The guanylin peptide family is involved with peptides secreted into the intestinal lumen related to water secretion and sodium absorption inhibition. Its discovery relates to the heat-stable enterotoxin Escherichia coli, a cause of diarrhea. Uroguanylin, encoded by the GUCA2B gene, is also a member of the guanylin peptide family, together with limphoguanylin. Missense SNP variations are, in most cases, neutral or have little effect on protein function, however, when these variations cause a change in the protein structure, this change could also result in a change in protein function. In this context, computational methods that predict the impact of missense SNPs on the structure of proteins are important to filter variations that could be studied in more detail using more refined methods, such as molecular dynamics simulations. In order to evaluate these structural alterations, molecular dynamics simulations are usually done in the range of nanoseconds, as longer simulations, which are more accurate, require a system with greater computational capacity. In this study, an analysis of the structural impact of missense SNPs present in the coding region of the GUCA2B gene was carried out using 50 SNPs as inputs, all found from the dbSNP database, which were evaluated by 16 in silico tools that predict the impact of variations on the protein structure, in order to select the convergent deleterious variations for further evaluation by molecular dynamics simulations with a time of 1 microsecond. Preliminary results obtained by simulations at 50 nanoseconds of duration suggest that 6 variations (Asp98His, Cys100Gly, Ala107Glu, Cys108Tyr, Gly110Ser and Gly110Asp) resulted in structural changes in the GUCA2B protein in terms of flexibility, stability and surface area accessible to the solvent. Moreover, the data reported here could lead to a better understanding of the structural and functional aspects of the uroguanylin peptide.

Identification and phylogenetic analysis of Cladosporium laccases

Ester Mota, Renata Guerra-sá

Universidade Federal de Ouro Preto

Abstract

Laccases are enzymes of multicopper oxidase family (benzenediol:oxygen oxidoreductase, EC 1.10.3.2) and has copper atoms working as a ligant. The major role of laccases in lignin and phenolic compound degradation has been evaluated in a large number of biotechnological applications such as dye degradation, bioremediation of some toxic chemical wastes. The aim of this work is detect the relationship among laccases of Cladosporium fulvum and laccases of other species as well determinate the putative cellular localization and protein structure. Amino acid sequences of laccases were recovered from JGI C. fulvum genome and a BLASTp was made to select laccase sequences. An analysis in Pfam was made to filter only laccase sequences that presented motifs of histidine to copper binding, characteristic of all laccases. 74 amino acid curate sequences of laccase were recovered from Swiss Prot and multiple alignments were performed with ClustalOmega and the phylogeny, using Kimura algorithm Neighbor Joining method, 1000 replicates. The 21 laccase sequences also were submitted to SignalP 4.1, to detect cellular positioning and were input in Swiss-Model to make protein modeling regarding 30% of identity, at the least. The phylogenetic tree related to domain sequence was more representative due to higher bootstraps and branching of plant laccases and fungi laccases, showed high level of conservation just closely to motifs. SignalP 4.1 analysis predicted that 11 laccases are extracellular, 8 intracellular and 2 transmembrane. SignalP 4.1 analysis predicted that 11 laccases are extracellular, 8 intracellular and 2 transmembrane. The identity of specific motifs in the amino acids sequences of the proteins showed a differentiation between plant and fungi enzymes. It was identified 5 laccase structures of 21 previous sequences, reinforcing the low degree of conservation among fungi laccases. Taken together, these results suggested that C. fulvum it is a potential source of new laccases which increases the biotechnological applications of this set of enzymes. Supported by: CAPES, FAPEMIG/VALE and UFOP

In silico study of detoxification protein Gluthatione S-transferase delta class from Anopheles darling

Marina Luiza Saraiva Möller, Ronaldo Correa da Silva, Nelson de Alencar, Rafael Sousa, Adonis de Melo Lima

Faculdade Integrada Brasil Amazônia

Abstract

A super family of proteins Glutathione S-transferase (GST's) is classified into several classes (Epsilon, Zeta, sigma, omega, delta and many others) acting as protection against oxidative stress and detoxification processes. In insects, delta and epsilon classes are responsible for the resistance to organochlorine insecticides such as Dichlorodiphenyldichloroethylene (DDT) via an elimination reaction through their cofactor gluthationa (GSH). This resistance in malaria vectors has been registered in ascending order since the 50s, when DDT began to be used on a large scale. The genome discovery of the major malaria vector in Amazon region (Anopheles darlingi) in 2013 made it possible a deeper study of resistance against insecticides in the locality. This work was conceived to presents a delta class protein of GST's from Anopheles darlingi and its conformational mechanism by bioinformatics perspective. For this, the aminoacid sequence of interest was taken from GenBank server (ID: ETN63518.1) and the three-dimensional structure was elucidated with the program Modeller 9:10, using a crystallographic protein from the Protein Data Bank server (ID: 1jlv) like a template, thus generating a model with 207 aminoacids, 5 alpha-helices and 4 beta-sheets, the canonical GST's presentation. The template and the model system was constructed and parameterized with the AMBER12 software package without ligand, using ff09SB force field. 50 nanoseconds of molecular dynamics were computed to evaluate the structural stability of model and template. The analysis of results was made done using behavior graphs of C-alphas and backbone atoms during simulation. A root mean square deviation (RMSD) and root mean square fluctuation (RMSF) plot showed the spatial conformational changes between atoms smaller than 3, which prove structural stability of template and model. Thus, based on the results, it is possible performed molecular dynamics with the GSH ligand, next step of work present here.

Prediction of affinity constant in the molecular docking using machine learning methods

Karla Machado, Laurent Dardenne, Leonardo Batista, Thais Gaudencio

Universidade Federal da Paraíba, Laboratório Nacional de Computação Científica, Universidade Federal da Paraíba, Universidade Federal da Paraíba

Abstract

The understanding of protein-ligand molecular recognition is one of the main aspects in structure-based design and discovery of new drugs. One key methodology is the docking of small molecules in protein binding sites with two aims: the search for the native ligand-protein conformation and the free energy calculation of this association, or its affinity constant prediction. To estimate the free energy values, the test set used in this work was composed by 50 protein-ligand complexes, which was extracted numeric characteristics considered important: Lennard-Jones' and electrostatic interaction energies values, ligand-receptor contact area associtaed with the solvent accessible surface, the presence of hydrogen bonds, and the number of the ligand rotatable bonds that were frozen in the process of docking. The estimated value of free energy was founded by using machine learning methodologies: Neural Networks, Linear Regression and Support Vector Machine (SVM), available in Weka and MATLAB softwares. For all, were evaluated the correlation and absolute error rates between expected and observed values. The tests were made with two normalization: linear and by standard deviation, since the attributes are in different numerical ranges. For both experiments, the absolute error rate ranged from 0.1992 to 1.1651 and the correlation rate ranged from 0.56 to 0.92. The SVM method obtained the best results. However, in MATLAB, tests were made with others architectures of neural networks, ranging the number of hidden neurons. The best results were found with the number of hidden neurons equal to 26 with correlation 0.95 for the test step. Under diversity aspects of the ligand, families, charges and among other features, the test set used in our work has a reasonable variability, therefore, at first it would be a good test set for the construction of a general empirical function (non-family or ligand dependent). However, higher error rates were estimated for complexes with lower and higher affinity. The number of protein-ligand complexes showing values of affinity constant in the range 1 to 3 and 7 to 11 are few, 17, in contrast with the higher number, 33, of complexes showing values in the range 3 to 7. Thereby, the SVM and Neural Network techniques showed the best results and is a good option for developing new receptor-ligand empirical free-energy functions. However, in future studies, it is necessary to increase the number and the diversity of receptor-ligand complexes to obtain more reliable empirical functions.

Comparative secretome and interactome analysis of pathogenic and non-pathogenic Trypanosomes

Renata Watanabe Costa, Ramon Oliveira Vidal, Fernando Antoneli Junior, Diana Bahia

UNIFESP, German Center for Neurodegenerative Diseases (DZNE), UNIFESP, UNIFESP/UFMG

Abstract

Trypanosomes are flagellate protozoa that inhabit various tissues of their hosts. Some of them can cause serious illness for humans such as Chagas disease (Trypanosoma cruzi) and Sleeping Sickness (T. brucei). On the other hand, some trypanosomes such as T. rangeli, which displays high genetic similarity with T. cruzi, are unable to induce pathogenesis in human. This intriguing fact has encouraged comparative studies between pathogenic and non-pathogenic species. For example, pathogenic species secrete proteins that can manipulate multiple host cell signaling pathways related to immune response and phagocytosis, favoring their invasion, survival and proliferation. However, there is a lack of comprehensive studies about specific secreted and transmembrane proteins of pathogenic trypanosomes. With this regard, bioinformatics tools can help to discover this set of proteins in order to identify their relationship to pathogenesis. In this study, we present an integrated computational pipeline for the analysis of secreted and transmembrane proteins and also preliminary findings about these proteins in some trypanosome species. The bioinformatics pipeline was mostly based on locally installed programs from the Center for Biological Sequence Analysis, such as SecretomeP, SignalP and TMHMM. We separated proteins that are exclusively secreted from proteins that are both secreted and transmembrane proteins, excluding those belonging to internal membranes. By using this approach, we were able to determine differences in secreted and transmembrane proteins between pathogenic (T. cruzi and T. brucei) and non-pathogenic (T. rangeli and T. evansi) species. In T. cruzi, we identified 658 secreted proteins; from this 195 are both secreted and transmembrane proteins in comparison with 492 and 269, respectively in T. rangeli. Among the secreted proteins identified in T. cruzi, we found several sialidases proteins, which were previously characterized by other studies and shown to be related to T. cruzi virulence. Regarding T. brucei, we found 101 secreted proteins in which 30 are both secreted and transmembrane proteins in comparison with 364 and 36, respectively, in T. evansi. The next step is to select some proteins expressed only in the pathogenic species and analyze their phylogeny, biological functions and interactions with host immune system proteins. The use of biological and computational methods will be critical for understanding the complexity of the host-parasite interaction and disclose biological mechanisms underlying Chagas Disease and Sleeping Sickness. These results will pave the way for a better understanding of their pathophysiology and ultimately leading to the identification of molecular targets for drug development.

Analysis of the accuracy of ASAProt (AUTOMATIC STRUCTURAL ANNOTATION OF PROTEINS)

Ana Larissa Gama Martins Alves, Caio Bulgarelli, Paulo Mascarello Bisch, Manuela Leal Da Silva

Instituto Nacional de Metrologia, Qualidade e Tecnologia (Inmetro), Universidade Federal do Rio de Janeiro - UFRJ

Abstract

Annotation is a methodology that seeks to infer the molecular function of proteins based on 3D structure and/or amino acid sequence. The basis for a correctly structural annotation for a protein include the sequential alignment and the choice of a good template. However, other tools can improve the annotation data such a rigorous validation parameters for models and the addition of the structural alignment. A software that uses simultaneously tools for sequential annotation and structural annotation has not been described in the literature yet. Therefore, a new structural annotation tool called ASAProt (Automatic Structural Annotation of Proteins) is under development integrating the portal MHOLline, generating specific models by comparative modeling and using a variety of software and databases for both annotation methods. Our goal in this study consists in manually analyze a group of selected proteins following the workflow of ASAProt. Furthermore, evaluate the efficacy of ASAProt through a comparison between the results obtained in the manual analysis and the results of the automatic method and determine the efficiency of the software. From the selected proteins, we obtained 10 models by the manual method and 8 models by the automatic method. A comparison between those models and templates were performed in order to apply the validation parameters for analysis. All the sequences compared presented a RMSD score under 8.14 sequences demonstrated a coverage higher than 70%. From the Ramachandran plot, all sequences that ran by the two methods, obtained more than 75% of their amino acids in the most favorable zones and less than 4% of their amino acids in unfavorable zones. R1 (>75%) and R4 (<4%) zones. Concluding, we can infer from the preliminary results that the validation parameters were performed successfully. After adjustments we will retest the two sequences that did not run by the automatic method. Other steps from the automation of the workflow still need evaluation.

Rama: A machine learning approach for ribosomal protein prediction in plants

Thales Francisco M. Carvalho, José Cleydson F. Silva, Elizabeth P. B. Fontes, Fabio R. Cerqueira

Universidade Federal de Viçosa

Abstract

Ribosomal proteins (RPs) play a fundamental role within all type of cells, as they are major components of ribosomes. Furthermore, these proteins are involved in various physiological and pathological processes. For instance, RPs have been demonstrated to trigger the tumor suppressor p53 pathway in response to ribosomal stress. In plants, RPs have been found to act in antiviral responses. These facts motivate advanced studies for the identification of unrevealed RPs. Current computational methods for functional genomics, such as InterProScan, employ comparative approaches that interconnect multiples databases of protein families. In this work, we propose a new in silico method for the prediction of RPs, termed Rama, based on machine learning (ML) techniques, with a particular interest in plants. We have realized that many unannotated proteins had great potential to be RPs. Furthermore, the classification of proteins with ML methods is much faster compared with methods that access multiple databases. To perform an effective classification, Rama uses a set of fundamental features of the amino acid side chain: is aromatic, is-negatively-charged, is-nonpolar-aliphatic, is-polar-uncharged, is-positively-charged, is-hydrophobic, in addition to molecular-mass, volume, and length. For our experiments, amino acid sequences of monocotyledons (Zea Mays and Oriza sativa), micotyledons (Arabidopsis thalina, Solanum licopersicum, and Glycine max), and phytoplancton (Ostreococus lucimarinus) were retrieved from Phytozome database v.10. Due to the similarity of RPs and histones, two training sets were created for each species, the first composed of ribosomal and non-ribosomal proteins, and the second composed of ribosomal and histone proteins. Three largely used ML algorithms were tested on these datasets: Multilayer Layer Perceptron, Randon Forest, and Support Vector Machines. Cross-validation and cross-species tests were performed to estimate the best algorithm and parameters for each species. As a result, Rama applies a two-step procedure to classify a set of proteins with unknown function as RPs: First, the models created from the datasets of RPs/non-RPs are applied; second, the proteins classified as RPs are given as input to the models created from the datasets of RPs/histones. Only the proteins considered as positives in these two classification filters are reported as RPs. Our results show that Rama could achieve an average accuracy, sensitivity, and specificity of 0.90, 0.82, and 0.92, respectively. Furthermore, our models classified several unannotated proteins as RPs with very high probability. Finally, the running time of Rama was approximately 600 times faster compared with InterProScan.

Development of a computational protocol to improve the affinity scoring of SVMPs inhibitors.

Raoni Souza, Natalia Fernandez, Rafaela Ferreira, Eladio Sanchez, Ronaldo Nagem, Adriano Pimenta, Rodrigo Ferreira, Francisco Schneider, Dimas Suárez

Universidade Federal de Minas Gerais, Universidad de Oviedo, Fundação Ezequiel Dias,

Abstract

Approximately 90% of accidents with venomous snakes in South America are caused by Bothrops species. Their venoms are characterized by a high proteolytic and hemorrhagic activity mainly caused by snake venom metalloproteinases (SVMPs), a group of toxins with a highly conserved catalytic site containing a zinc(II) ion. Drug-like inhibitors of these toxins could represent a complementary procedure to serum therapy and help to develop more effective treatment for the local effects. These inhibitors could be discovered through the use of virtual screening (VS) followed by molecular dynamics (MD) simulations and end-point free-energy methodologies to refine the relative binding affinity of potential inhibitors. However, selection of the settings for the preliminary docking calculations, MD simulations and end-point free-energy calculations is a challenging task due to the presence of the Zn(II) ion and the plasticity of the SVMP binding pocket . In this work we report a computational protocol that could be used to improve the affinity predictions for SVMP inhibitors using a docking protocol specially developed for metalloproteinases, followed by geometry optimization with hybrid quantum mechanical/molecular mechanics (QM/MM) calculations, MD simulations using selected docking poses, QM/MM &PBSA (poisson-boltzmann surface area) calculations on MD snapshots as well as conformational entropy estimations. To test this protocol, we studied six broad-spectrum metalloprotease inhibitors, bearing a hydroxamic group, with in vitro activity against the hemorrhagic toxin atroxlysin-I of Bothrops atrox and the non-hemorrhagic toxin leucurolysin-a of Bothrops leucurus, which were selected as SVMP models. The computed relative binding energies were compared with the experimental binding affinity of the inhibitors inferred from kinetic assays. We also characterized the enzyme-inhibitor interactions, identifying the major differences between the two toxins. Although the correlation between the computational and experimental data is not perfect, all the examined compounds were evaluated as potent inhibitors with similar affinity for the toxins in agreement with experimental data, suggesting thus that this protocol could be used to improve virtual screening (VS) results. Hence, the next step would be to automate this protocol aiming to use it for few tens of compounds selected in VS studies.

DEVELOPMENT OF COMPUTATIONAL TOOLS FOR BIOLOGICAL DATA ANALYSIS AND PROTEINS IDENTIFICATION FROM METAGENOME SAMPLES

Rafael Nicolay Baptista da Silva, Manuela Leal da Silva

Instituto Nacional de Metrologia, Qualidade e Tecnologia, Instituto Nacional de Metrologia, Qualidade e Tecnologia, Grupo de Pesquisa em Biologia Computacional – DIMAV – INMETRO, Programa de Pós-Graduação em Biotecnologia – INMETRO

Abstract

The rearrangement of different methodologies for data analysis and curation of omics raw data are essential for a general distribution and identification of biological patterns. The employment of annotation strategies, as sequential and structural, can improve the biological relevance and quality of a metagenome sample. In this study, our goal consists in presenting a sequential annotation strategy for specific data identification by the rearrangement of different methodologies described in the literature. As input for our research, we use protein domains characterized as Glycosyl Hydrolase family 2, Carboxyl Esterase family 6, Glycosyl Hydrolase family 10 and Carboxyl Esterase family 1. Further, we perform a search in order to identify patterns using biological libraries containing sequence fragments from two different metagenome samples from: Achatina fulica and Bradypus variegatus. As methodology, we apply local alignment techniques for the identification of biological patterns, as identity and similarity, using statistic and probabilistic software as BLAST, HMMER. Forward, for each match with our input sequences, we perform a new search using the best sequence fragments for identify annotated data. On this step, we use local and global alignment techniques with BLAST, COBALT and ClustalW2 software from a web server platform against the NCBI databank. In order to refine and corroborate our results, we perform phylogenetic reconstruction using the results from the previous analysis and applying distance evolution and maximum-likelihood methods. We execute a reverse search procedure using an annotated sequence against the biological samples in order to identify sequence fragments. At last, we perform a reference-based assembly and annotation of a new protein, on both metagenomes sample from the fragments, using the annotated protein as the reference sequence. As preliminary results, we obtained two potential sequences with a high rate of identity characterized as endo-1,4-β-xylanase and sialic acid-9-O-acetylsterase from the Bacteroides sp. . As conclusions, the methodology employed was efficient for the specific protein identification. As perspectives, we will perform the analysis using other omics sample. Further, we will automatize this methodology. Supported by: CAPES and CNPq.

Use of computational methods for the evaluation of the structural and functional impact of missense SNPs present in the CYP2D6 gene

Leonardo Fialho, William Porto, Sérgio Alencar

Universidade Católica de Brasília

Abstract

Belonging to the family of cytochrome P450 system in the oxidative metabolism of drugs, the CYP2D6 gene has great functional importance. This gene is responsible for encoding the CYP2D6 enzyme, which is able to metabolize approximately 60% of the drugs passing the liver, and may be directly responsible for drug absorption. There are currently deposited in the dbSNP database a total of 191 missense SNPs present in the CYP2D6 gene. Several of these variations have been associated with changes in drug metabolism. Our hypothesis is that missense SNPs present in this gene could cause significant changes in the structure of the CYP2D6 enzyme and, therefore, could lead to functional changes. Currently there are no published works involving a comprehensive study of the structural impact of missense SNPs in this enzyme using computational methods such as molecular dynamics and molecular docking. From the 131 missense SNPs obtained from the NCBI's Variation Viewer database, in silico analysis was carried out, using a total of 16 computational tools (SIFT, Provean, Mutation Assessor, Panther, SNAP, PhD-SNP, Suspect, PolyPhen, EFIN, Site Directed Mutator, Fold-X, PoPMuSiC, Condel, Meta-SNP, PON-P2 and PredictSNP) divided into 4 distinct groups, according to their prediction algorithms. Then, we filtered all missense SNPs that were classified as deleterious by at least three tools in each of the four different groups, and called these as convergen deleterious predicted SNPs. Following this procedure, we identified 23 convergent deleterious predicted SNPs. From this result, we will continue analyzing the impact of the SNPs through molecular dynamics simulations and molecular docking.

Bioinformatics as a valuable tool in identifying forms of PDC-109 in the cryopreserved seminal plasma of Bos taurus indicus bulls

Marcos Jorge Magalhães-Junior, Denise Silva Okano, Thaís Ferreira dos Santos, Leonardo Franco Martins, Renato Lima Senra, Paulo Roberto Gomes Pereira, Alessandra Faria-Campos, Sérgio Vale Aguiar Campos, José Domingos Guimarães, Maria Cristina Baracat-Pereira

Universidade Federal de Viçosa, Universidade Federal de Viçosa, Universidade Federal de Viçosa, Universidade Federal de Viçosa, Universidade Federal de Viçosa, Universidade Federal de Viçosa, Universidade Federal de Minas Gerais, Universidade Federal de Minas Gerais, Universidade Federal de Viçosa, Universidade Federal de Viçosa

Abstract

PDC-109, a bovine seminal plasma (BSP) protein composed by BSP-A1 and BSP-A2, is involved in fertilization and semen cryopreservation processes. PDC-109 is described to be both beneficial and detrimental to sperm but its participation in freezability events is not completely understood. By two-dimensional electrophoresis (2-DE), we have evidences that more than two forms of PDC-109 are present in seminal plasma of fertile Nelore bulls showing differences in semen freezability. Not all animals presented similar abundance of each PDC-109 form. However, the small number of analyzable fragments generated by trypsinolysis hampers the significant identification of PDC-109 by frequently used proteomics tools; large number of cleavage sites for trypsin is present in the sequence of PDC-109. The goal of this work was to use computational tools aiming the significant identification of the forms of PDC-109 in the cryopreserved seminal plasma of Bos taurus indicus bulls. Proteomic analysis by 2-DE has evidenced differential abundance of seven protein spots for frozen/thawed seminal plasma among the nine bulls. Characteristics of the seven spots (15 to 17 kDa, pI 4.6 to 5.8) were in accordance with PDC-109, however Mascot has not identified PDC-109 ($P<0.05$). The sequence of PDC-109 was then theoretical cleaved by trypsin using PeptideCutte (with addition of 57.05 Da for each alkylated cysteine). The five theoretical ions (kDa) were detected in the MS spectra, for the four spots of interest, and the different ions in MS/MS spectra were manually sequenced by de novo procedure. The obtained sequences coincided with the theoretical ions. The commercial softwares Biotools (v.3.2, Bruker Daltonics) and Mascot Daemon (v.2.3.2, Matrix Science), using a protein database for the bovidae family (from the UniProt database), endorsed the identification of PDC-109 ($P<0.05$) in the four spots for all bulls. The Bovine Genome Database confirmed the proteins. One spot (15.52 ± 0.53 kDa, pI 5.78 ± 0.12), must be highlighted by its high abundance in bulls with low semen freezability and its absence in bulls presenting high semen freezability. This is the first evidence that more than two forms of PDC-109 are found in seminal plasma of Nelore bulls, and that not all animals present similar abundance of each PDC-109 form. The available computational tools have aided proteomics in the identification of the PDC-109 forms. Supported by FAPEMIG, CNPq, CAPES and FINEP. Thanks to Agropecuária CFM Ltda/SP, NuBioMol/UFV and BIOAGRO/UFV/Viçosa-MG, Brazil.

Enthalpic and entropic factors as determinants for the different pattern of affinity of galantamine and derivatives by the human acetylcholinesterase - A molecular dynamic study

Rafael Eduardo Oliveira Rocha, Leonardo Henrique França de Lima

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

Abstract

A recurring target for mitigating the harmful effects of Alzheimer's disease are acetylcholinesterase, a globular protein present in the cell membrane of neurons. The current approved drugs seek to inhibit effectively its enzymatic action of acetylcholine degradation, increasing the duration of the synaptic signal and consequently, reducing the symptoms from neural degeneration. One of the medications, also named as the most promising of his generation, is galanthamine. Her has a three-ring structure, rich in polar atoms, allows her to maintain a network of hydrophilic interactions and stacking with specific residues of the active site of the enzyme. Experimental data show that two galanthaminics derivatives have inhibitory activity distinct. The lycoramine, whose configurational difference with galanthamine is determined by substituting a double bond for a single in one of the rings, has substantial decline in inhibitory activity. In return, the sanguinine, with substitution of a methyl group for a hydroxyl in galanthaminic structure, has increased inhibitory activity. This study aims to understand the thermodynamic standards governing these inhibitory differences by molecular dynamics simulations of galanthamine, sanguinine, lycoramine and a hybrid of sanguinine and lycoramine (molecule wherein the two structural differences in sanguinine and lycoramine were added concurrently to the structure galanthaminic) in the active site of the protein. Simulations done for 20 ns in AMBER force field have demonstrated that replacement of the hydroxyl, in case of sanguinine and hybrid, enable these molecules to obtain an advantage in terms of electrostatic interactions with the protein in relation to the remaining two. However, in the case of the hybrid, the loss of the methyl group severely diminishes its ability to maintain interactions of Van der Waals, which does not occur for sanguinine, because keeping the SP2 geometry of the ring, the pattern of hydrophobic interactions is optimized, as well as in lycoramine. The galantamine maintained at mediocrity compared to other ligands. The last, however, although it is present in an intermediate position in the above parameters, demonstrates a high contribution of conformational microstates when compared to other ligands, as the analysis of its total energy throughout the simulation for internal RMSD reveals two well defined cluster, while for the other ligands, only one is discernible (except lycoramine, but the second observable microstate is very little accessed). It is hoped that with better understanding of thermodinamics contributions of changes in galanthaminic structure allows for more rational design of inhibitors for acetylcholinesterase.

Over the Thermostabilization Mechanisms of Two Punctual Mutations in the Bacillus polymyxa β-glucosidase A – A Molecular Dynamics/Docking Study.

Tiago Silva Almeida, Rafael Eduardo Oliveira Rocha, Leonardo Henrique França de Lima

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

Abstract

The industrial production of the named lignocellulosic ethanol (or "second generation" bio-ethanol) presents itself as a renewable, "clean" and economically viable alternative to the usual fossil fuels. Such production is dependent of the integrated activity of a range of microbial enzymes, among which the β-glucosidases control a limitant step. In order that the global process is economically and thecnically feasible, it is interesting that such limitant enzymes present themselves high specific activity and that they resist to extremes conditions on the industrial reactors, both conditions that can be reached using enzymes of higher thermostability. For the Bacillus polymyxa mesophilic β-glucosidase A, there is a set of punctual mutations described in literature that enhance considerably the enzyme thermostability, many of which with their molecular mechanisms still not completely understood. In this study, we have used molecular modelling and molecular dynamics simulations at 300 and 363 K to probe such stabilization mechanisms for two punctual thermostable mutants – E96K (TR1) and M416I (TR2) – compared to the native enzyme. The simulation results point that a local new salt bond betwen the mutated E96K residue and D28 promotes a higher stabilization of the loop 2 dynamics in 1TR1, limitting collective movements that, in the native protein, promotes distortions at the active site at 363 K. For TR2, a pretty more "rigid" hydrophobic packing of I416 compared to the original M416 seems to promote a dynamic stabilization of the N-terminal β-hairpin what, in turn, contributes both to compact relatively the structure of the protein as a whole as, again, to reduce collective movements that lead to the distortion of the active site at 363 K. Molecular docking assays of the substrate cellobiose to the active site of clusterized structures from each simulation show a weighted $\Delta\Delta G_{\text{Docking}}$ values from 300 to 363 K for the native, TR1 and TR2 enzymes respectivelly of 0.8, -0.3 and 0.06 kcal.mol^{-1}. In this way, the subtle conformational and dynamic changes observed in the mutated systems seem really to preserve a higher affinity of the enzyme by the substrate at high temperature compared to the native one. We expect that such results promote a higher understanding of the mechanisms by which specific mutations improve the thermostability of β-glucosidases, allowing future works of protein engineering to generate more industrially adaptated enzymes.

Integrating transcriptomics, proteomics and physiology scales in sugarcane roots

Amanda Rusiska Piovezani, Fabrício Martins Lopes, Marcos Silveira Buckeridge

University of São Paulo, Federal University of Technology, University of São Paulo

Abstract

Systems can be defined as structures that are formed by elements that interact giving rise to a meaning of function. In biology, understanding how systems function requires knowing what elements form systems, how these elements interact with each other in order to give rise to the emerging biological function of the system. Biological systems perform their functions by integrating actions at different scales that lead to the emerging overall function of the system. It is now clear that every biological process will be funded on mechanisms related to the scales of genome, transcriptome, proteome, metabolome and phenome. In the literature, these scales are usually studied separately, i.e. the biological function is usually explained by the individual mechanisms associated to one or two of the scales . The present work tries to evaluate all the scales mentioned above in an integrated form so that a much wider systems approach can be visualized. Here we used as a model a process of modification of cell walls within the roots of sugarcane. During development of the roots, the tip of the root contains cells that are dividing and producing cell differentiation through a complex network of signals that lead to cell expansion and elongation with the concomitant formation of the vascular system. At the same time, the cortex undertakes changes in cell walls that resemble degradation. In both cases, development is associated with programmed cell death and cell wall modifications that lead to the emerging functions of nutrient and water transportation and oxygen access to the living cells through a structure named aerenchyma that forms in the cortex. We have concurrently analyzed transcriptomics, quantitative proteomics, metabolomics and phenomics of the developmental process of sugarcane roots. Integration of the data afforded the discovery of a collection of key genes related to the mechanisms associated to the systems mentioned above. The integrated view of events occurring at the same time was collapsed to single map that describes wall-related events during aerenchyma formation system at different scales at the same time. After having all these information analyzed, we are now putting together a computational tool that can help visualization of the whole system. We expect that with this new tool, it would be possible to design modulations of the whole system in much more precise and reliable ways so that plant engineering for cell wall hydrolysis can help development of technologies for bioenergy production from grasses.

Functional profile of a copper mine tailings dam

Laura Leite, Julliane Medeiros, Sara Cuadros-Orellana, Gabriel Fernandes, Guilherme Oliveira

Universidade Federal de Minas Gerais, CPqRR/Fiocruz, Instituto Tecnológico Vale

Abstract

Metal-contaminated freshwater habitats exhibit an extremely complex and well adapted community. These microorganisms have enzymes with several biotechnological applications that play an essential role in environmental biogeochemical cycling, transport and storing of contaminants (metal bioavailability). Here we obtained an overview of the taxonomic and functional microbial diversity of tailings dam in a Brazilian Amazon mining area that received more than 90 million tons of chalcopyrite mining waste. We collected 3 liters of surface water (1M), water from 15 meters depth (15M), and sediment (SED). The prokaryotic biomass from water samples was concentrated with a sequential membrane filtration system. DNA extraction was performed for the membranes of 0.22m pore size and sediment. The shotgun metagenomes were sequenced on Ion Torrent platforms, a total of 16 samples were produced, each one with 1Gbp. A taxonomic profile was obtained by comparison all reads against the NR database. BLAST file were imported into MEGAN to calculate a taxonomic classification based on the lowest common ancestor method. This analysis revealed a complex microbial community with a dominance of Porphyrobacter, Gemmatimonas and Ilumatobacter at 1M, Anaeromyxobacter, Burkholderia and Bacillus at 15M, Ilumatobacter, Sediminibacterium and Burkholderia in the sediment. Contigs >500bp (1M: 46792, 15M: 28751, SED: 9239) were obtained with SPADES assembler. Coding sequences were predicted with MetaGeneMark (1M: 125480, 15M: 59850, SED: 14676). Proteins were extracted by one in house script and classified into Uniprot protein families using blastp. KEGG orthology groups were identified with blastx and mapped to KEGG metabolic pathway modules using Ipath2. Functional reconstruction revealed a diverse set of genes for ammonium assimilation and ammonification in the water (1M, 15M), maybe derived from leached silicate minerals. As expected, pathways associated with sulfur cycling (sulfur-oxidation and sulphate-reduction) were found in SED samples. Functional annotation unveiled a diversity of metal resistance genes in all layers, suggesting that the prokaryotic community is adapted to metal contamination.

Prediction of co-regulation between microRNAs and transcription factors: a Bayesian network model for the study of regulatory associations

Vinicius Chagas, Mauro Castro

Universidade Federal do Paraná

Abstract

Transcription Factors (TFs) and microRNAs (miRNAs) are important cellular regulators that act on a variety of cellular processes such as proliferation, differentiation and apoptosis. Several studies have shown functional associations between TFs and miRNAs in cancer phenotypes. These associations can potentially establish feedbacks and feedforward loops that might modulate the downstream effects on the common target genes. While co-regulation between miRNAs and TFs is a plausible biological phenomena, there still exists a large methodological gap between the individual description of each regulator and a fully systems integration. Here we present a new computational approach able to infer the probabilistic dependencies between TFs and miRNAs using mutual information and Bayesian inference. The algorithm is implemented in R language and extends functionalities from the RTN and RedeR packages, available from the Bioconductor. We demonstrate the workflow using data from 126 bladder cancer samples selected from the TCGA platform with RNA-seq data available for both RNA classes. The resulting regulatory networks comprise 1058 regulons (819 centered on TFs and 239 on mature miRNAs). Using the bnlearn package to derive co-regulatory information we identify 348 associations with probability > 0.95, and 4 TF-miRNA potential interactions: TCF21 miR-29c-3p, BATF miR-29c-3p, ZNF20 miR-29c-5p and MEOX1 miR-150-5p. Our results are consistent with previous studies describing these regulatory elements in cellular development process (proliferation, differentiation or apoptosis) and several cancer phenotypes. Also mir-29 family members have been associated with bladder cancer, and tumor suppression in colon and breast cancers, suggesting that the predicted TF-miRNA interactions are potentially interesting markers to follow up in subsequent studies.

PREDICTION OF PROTEIN INTERACTION NETWORKS BASED ON STRUCTURAL INFORMATION OF PREDICTED PROTEINS IN GENOMES OF LEISHMANIA.

Crhisllane Rafaele dos Santos Vasconcelos, Thais Helena Chaves Batista Antonio Mauro Rezende

Universidade Federal de Pernambuco, Fundação Oswaldo Cruz (Fiocruz) - Pernambuco - Centro de Pesquisas Aggeu Magalhães (CPqAM)

Abstract

In according to the World Health Organization, 1-2 million new cases of leishmaniasis occur each year. The available drugs for treatment have serious drawbacks, and no effective vaccine has been developed. Thus, applications of systemic approaches to discovery of new drug/vaccine targets are needed. One of these approaches is the study of protein interaction networks. Therefore, the main goal of this work is to model protein interaction networks for Leishmania braziliensis and Leishmania infantum, two important causing agents of leishmaniasis, from their predicted proteomes based on structural information. In order to reach this aim, protein sequences of both proteomes were obtained from the TritrypDB. With the BLAST tool, these sequences were aligned against the PDB proteins, and templates were selected according to their identity and the alignment coverage. The leishmania protein and PDB template were then used as input for MODELLER package. Proteins, that templates were not found by BLAST search, were submitted to MHOLline, ModPipe and Phyre server. All generated models were assessed by PROCHECK and normalized DOPE score generated by MODELLER. A total of 8357 and 8239 proteins of L. braziliensis and L. infantum, respectively, were obtained from TritrypDB. From these, 2212 and 2225 proteins were modelling. These, 1677 and 1683 models showed less than 1% of their torsion angles in not allowed regions in Ramanchandran Plot and values of normalized DOPE lower than -0.5. All qualified models will be submitted to molecular docking using Megadock and Patchdock tools, and a template-based protein complex structure prediction tool called Prism, for construction of protein interaction networks.

FUNCTIONAL ANALYSIS OF PROTEIN NETWORKS FROM Aedes aegypti

André Luiz Molan, Carine Spenassatto Dreyer, Jayme Augusto de Souza Neto, José Luiz Rybarczyk-Filho

Instituto de Biociências de Botucatu - UNESP, Instituto de Biotecnologia de Botucatu - UNESP

Abstract

Dengue is an arbovirus whose number of people affected by the disease in Brazil has reached high numbers in recent years. Its main vector is the Aedes aegypti mosquito,which is found in tropical and subtropical regions of the planet. The anthropophilic haematophagic habit, own behavioral characteristics and rapid development, makes it a great dengue transmitter. Numerous studies have focused on a specific treatment and the development of a vaccine to prevent infection. The application of bioinformatic tools may help to understand the correlation between host and parasite relationship. Herein we analized RNAseq data from infected and not infected mosquitoes with dengue virus serotype 4 through the assembly of protein-protein interaction networks (ppi). We start from a differentially expressed gene pool obtained from RNA-seq experiments in Lights HiScanTMSQ platform. The data were divided into four parts (I, II, III and IV), comprising populations of mosquitoes infected and not infected with dengue virus serotype 4 in two different regions (Botucatu-SP and Neópolis-SE). The repository of protein-protein interactions STRING, was used to built four networks (S1, S2, S3 and S4) and perform the enrichment of the data to find all elements associated into the biological processes (BP) and molecular functions (MF). Finally, using the ViaComplex software, we projected the expression data and ontologies on the networks, to visualize the distribution of proteins and their respective ontologies.Taking into account just processes and functions with p-values equal or less than 10-3, 7 BP and 2 MF were identified in S1 network, 31 BP and 6 MF in S2, 15 BP and 10 MF in S3 and, in S4 network, 92 BP and 35 MF. Additionally, it was observed that all the differentially expressed genes in III and IV were present in the networks S3 and S4 respectively, unlike I and II, whose percentage was below 60

ARCoBALeno: an application for coloring biological pathways by ancestrality or gene function

Carlos A. X. Gonçalves, José M. Ortega

UFMG, UFMG

Abstract

Biological pathways are often used to visually depict the interactions established between gene products related to a biological process. Previous works have used algorithms or orthologues databases to determine the LCA (lowest common ancestor) of the gene products on a biological pathway, thus allowing for the study of that pathway evolution along the history of life. One way to visualize this information is by coloring the elements by their ancestrality, something that can be quite time demanding on larger pathways when manually performed. We developed ARCoBALeno (Application for Rapidly Coloring Biological pathways by Ancestrality or Last molecular function), a software to perform automated coloring on PathVisio XML pathways to easily allow for the study of their ancestrality. ARCoBALeno extracts the gene names on the graph's data nodes, determines the UniProt identifier for those gene names and unravel their ancestrality by calling SeedServer, a software to identify clusters of orthologues. For each level on the NCBI Taxonomy lineage for the inputted taxon, an output XML is generated by ARCoBALeno containing the gene products colored up to that point by their ancestrality. Thus, when observed in succession, these snapshots reveal the evolution of that biological pathway. ARCoBALeno also has access to a local database containing ancestrality analyses performed on the Gene Ontology (GO) base terms. Using this data, ARCoBALeno can query the GO molecular function annotation for the gene names on a pathway and attribute to each element the LCA color of the "most recently acquired function", which allows for studies regarding gene plasticity within a biological process. ARCoBALeno is fully compatible with the large library of pathways available on WikiPathways, since they can be downloaded as PathVisio files. An execution pilot was performed using the ATM Signaling Pathway deposited on WikiPathways (entry WP2516). The ancestrality analysis of the generated output reveals that a single gene on this pathway already existed by the origin the cellular organisms, probably with unrelated functions; the core of this process is clearly born at the base of the eukaryotes, where ATM itself arises, and from then the pathway continuously acquires new functionalities alongside the evolution of the human lineage. ARCoBALeno is available for public use as a web application at biodados.icb.ufmg.br. Funding support: FAPEMIG, CAPES

A method to modify molecular signaling networks through examination of interactome databases

Lulu Wu, Marcelo Reis, Vincent Noël, Hugo Armelin, Junior Barrera

Instituto de Matemática e Estatística, Universidade de São Paulo, LETA/CeTICS, Instituto Butantan Instituto de Química, Universidade de São Paulo

Abstract

Signal transduction is the primary means by which cells respond to external signals from their environment and coordinate complex cellular changes. The study of molecular signaling networks aims to understand the operation of each process of cellular signal transduction; in such studies, the usage of mathematical models to simulate the kinetics of chemical reactions that describe a given signaling network allows us to generate testable predictions of the cellular processes. To assist these modeling and simulation procedures, the Group of Computational Biology and Bioinformatics at Instituto Butantan recently introduced the SigNetSim e-Science framework. This framework makes the description of molecular signaling networks through a set of chemical reactions, which are mapped into a system of ordinary differential equations; after that, the parameters of this system are adjusted through curve-fitting optimization and the simulation result is evaluated. However, this framework has no systematic approach to test whether modifications in the signaling network (e.g. inclusion or removal of chemical reactions) enhance the fit of the network model simulation to the observed data. Hence, we present in this work a method to systematically modify molecular signaling networks and assess if a given modification improves its respective kinetic model. This method relies on the usage of interactome databases to provide a set of candidate chemical species to be included in a given signaling network. We developed a component within the SigNetSim framework to test different hypotheses (i.e. signaling network modifications), using a greedy strategy of chemical species inclusion. To evaluate the implemented component, we used KEGG as the interactome database and a MAPK/PI3K/Akt pathways model as case study. Initial experimental results showed this method is able to recover to some extent the case study signaling network from random subnetworks. Therefore, we believe that this strategy is promising to improve the mathematical modeling of molecular signaling networks.

VISUALIZATION OF BIOMOLECULAR NETWORKS USING FORCE-BASED LAYOUT IN A CELL

Henry Heberle, Hugo H. Slepicka, Guilherme P. Telles, Rosane Minghim, Gabriela V. Meirelles

Universidade de São Paulo, Brookhaven National Laboratory, Universidade Estadual de Campinas, Universidade de São Paulo, Centro Nacional de Pesquisa em Energia e Materiais

Abstract

With the advent of "omics" science, analyses performed from the screening of a wide range of physical, genetic and chemical-genetic interactions have brought new perspectives in contemporary biology, as they provide new clues in protein/gene function, in the organization of biological pathways and in the validation of therapeutic targets. Biomolecular interaction networks, or graphs, are simple abstract representations where the components of a cell (e.g. genes, proteins, metabolites, miRNAs, etc.) are represented by nodes and their interactions are represented by edges. An appropriate visualization of the data is crucial for understanding such networks, particularly regarding high-throughput analysis, since pathways are related to specific functions that occur in specific regions of the cell. The force-based layout is an important technique to draw networks according to their topologies. Dividing the networks into cellular compartments helps to quickly identify where network elements are located and more specifically concentrated. Currently, only a few tools provide the capability of visually organizing networks by cellular compartments, but cannot handle big and dense networks because of limitations of the grid layout approach. Here we propose CellNetVis, a web tool to visualize biological networks in the XGMML format considering the GO cellular components annotations. A force-based layout is applied to the network and the nodes are constrained in their own cellular component. To minimize the computationally expensive feature of the force-based layout, we developed an algorithm to draw the cell diagram based on circles, considering the position of the node relative to the center of the cellular component and their boundaries. The colors of the diagram were optimized and a constraint that prevents nodes to overlap was also added. Interactive features of the visualization tool allow the user to identify specific nodes, their interaction partners and values stored in their attributes. Network topology measures, edge bundling, selection of nodes, highlight of neighbors, display of labels, counting of nodes per cellular component, drag and drop of nodes and cellular components and the possibility to change the attribute related to the nodes colors in the network were implemented. Our bioinformatics tool was written in HTML and JavaScript and is capable of displaying information related to complex networks, nodes and edges as well as their relations with cell partitions. It is currently been validated on datasets generated from different organisms annotated with specific GO cellular components, which are displayed in the diagram accordingly.

Visual comparison of annotated biomolecular networks using all-in-one approach

Henry Heberle, Gabriela Vaz Meirelles, Bianca Alves Pauletti, Adriana Franco Paes Leme, Guilherme Pimentel Telles, Rosane Minghim

USP, CNPEM, CNPEM, CNPEM, UNICAMP, USP

Abstract

Networks that store connectivity information and node feature information can represent biological systems. In the context of molecular biology, these nodes may represent proteins, metabolites and other types of molecules. Each molecule is annotated and stored in databases, such as Gene Ontology. A visual comparison of networks requires tools that allow the user to identify differences and similarities between nodes attributes as well as between their known interactions (links). We are developing a technique to facilitate the comparison of these annotated biological networks, striving to maintain in the process the visualization of the network connectivity. The tool allows the comparison of up to six networks using the all-in-one approach. The user personalizes the final network according to the chosen set operations that will be applied on links, nodes, and on nodes attributes. Our current prototype shows the final network to the user, where exploratory functions are available. Unlike most known tools for visualizing biological networks, it allows the creation of networks whose attributes are derived from the original networks through operations of union, intersection and unique values. A visual comparison of the networks is achieved by visualizing the outcome of such joint operations. The comparison of nodes attributes can be also performed using Venn diagrams. To assist this type of comparison, the InteractiVenn technique was integrated to the system, in which the user can interact with Venn diagrams, performing union operations between sets and diagram shapes. This diagram union feature differs from other tools available for creating Venn diagrams. With these set of functionalities, users manage to compare networks from different perspectives. To exemplify and test our prototype, case studies are being carried out.

A network-based model for the study of regulatory genes in genome stability pathways

Marthin Borba, Mauro Castro

UFPR

Abstract

A biological phenomenon can be represented as interaction networks comprising several classes of molecules and processes. Metabolites, genes and proteins are examples of these molecules and the functional characterization of the interactions among them is still a major challenge. Here we present a systems biology framework to investigate the genome stability in normal and cancer cells. Using protein-protein interaction networks mapped to the human genome maintenance mechanisms we explore the regulatory elements that might explain the dysfunctional phenotype and the intrinsic instability observed in the chromosomes of many cancer subtypes. The network model was constructed from an initial survey of 1105 genes involved in several genome stability pathways, including apoptosis, inflammation, DNA damage and cell cycle. Genes are annotated according to their occurrence in each pathway, and those listed in more than one have been classified as overlapping genes. To demonstrate the main features of this network model we have used gene expression data from colorectal cancer (CRC) obtained from public repositories. We show that the network activity is altered in CRC compared to normal samples, especially in those samples linked to loss of genomic stability. This result is consistent with previous studies that demonstrate the high rate of genetic changes observed in unstable CRC subtypes. We anticipate that our network model will help to explore and better characterize regulatory genes associated with key processes that might lead to the genome instability in cancer cells. The statistical methods and algorithms developed in this study will be freely available in an R package.

Entropy of Network Information Flux in Glioblastoma Multiforme

Luis Henrique Trentin de Souza, Alfeu Zanotto-Filho

UNIVERSIDADE FEDERAL DE MINAS GERAIS, UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

Abstract

Glioblastoma Multiform (GBM) possesses high levels of genomic instability, which consequently promotes the formation of a variety of highly proliferative, invasive and chemoresistant cell phenotypes. Such a genomic instability in tumors potentially impact the "stoichiometric" balance between functionally related proteins, mostly due to changes in transcriptional activity with these entities. Analyzing gene expression correlations under a protein interaction network perspective allow studying the informational flux probability between linked proteins. It means that linked proteins with positive correlation are more prone to information flux than negatively correlated links. The local network entropy (LNE) is an index which measures the "disorder" of the informational flux between a protein and its neighbors. Increased local network entropy was recently shown as a systemic hallmark of diverse tumors. However, to best of our knowledge, there are no studies investigating the significance of local network entropy changes in brain tumors thus far. We believe that understanding the network informational flux, and the biological processes affected in this context, can bring new insights on the pathobiology and drugable pathways in GBM. In this study, we aimed to investigate; i) whether the local network entropy of GBM differs from normal brain tissues; ii) which genes displayed increases in entropy and what biological pathways or processes they are involved in; iii) if the identified biological processes/pathways carry differentially expressed genes and; iv) whether some of the differentially expressed belonging to highly entropic pathways are correlated with GBM patients survival. In view of such a aims, our results showed that: i) GBMs showed a significant increase in local network entropy values when compared to non-tumor brain tissues; ii) genes with high entropy played a role in 28 biological processes potentially related to GBM physiopathology; iii) Several genes with the identified pathways were found overexpressed or down-regulated in tumor versus normal brain tissues and; iv) amidst them, the expressions of PAK6, PLCB1, MAPK8, CDK6 e MYD88 predicted better prognosis, while overexpression of two Calcium/Calmodulin Kinase isoforms (CAMK2A e CAMK2B) were correlated to poor prognosis; an effect only observed in patients younger (<50 years-old) at the age of diagnosis. In summary, this study shows that local network entropy in combination with pathway enrichment analysis are a useful strategy to improve our knowledge on the biological alterations as well as genes relevant to prognosis in GBM under a systems biology perspective.

Applications of graph theory for verify the family relationship for genetic evaluations through the Animal Model

Pedro Bittencourt, Fernanda Almeida, Wagner Arbex

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

Abstract

Animal genetic improvement is an usual method for increasing the productivity of the herd. Therefore, selecting animals that are potentially b̈etter̈than their peers and/or contemporary is important because the descendants of these animals will have enhanced characteristics. Genetic evaluations are used in animal genetic improvement programs to predict the potential genetic value of the animals and their PTAs (predicted transmitting abilities); however, to obtain good results, one needs as much information as possible about the individuals, their relatives, and their ancestors to get the most accurate data. The Animal Model is a computational implementation of genetic evaluations which considers significantly the information about the kinship of animals. The Animal Model considers the data of each evaluated animal, as well as other animals that have the same relationship to it. Some of its best-known instances, such as MTDFREML, require recoding identification of all animals that are in the database, recoding them from the fathers, then the mothers, and finally the sons. By recoding animals in this way, it is possible to identify and, when appropriate, correct any existing irregularities in the set of data, such as the same identification being used for different animals, which can also cause the inconsistency of an animal being its own ancestor. Thus, a graph can be built, considering (i) the individuals and their family relationships as vertexes and edges, respectively, and (ii) the graph concepts and directed graph, the path, circuit, and cycle, and the connected/disconnected graphs and strongly connected graphs. Because the database might be extensive, one should generate many connected graphs, or better, a graph with connected components that resemble kinship trees between animals. The connected components are checked and in the observation of a cycle - a strongly connected sub-graph - it means that some individual of this sub-graph is its own ancestor. Supported by: CAPES, CNPq, Embrapa, Embrapa Dairy Cattle, FAPEMIG and UFJF.

Using network analysis to probe emergent properties of physiological behavior of whole organisms

Vinícius Jardim Carvalho, Suzana de Siqueira Santos, Amanda Rusiska Piovezani, Amanda Pereira De Souza, André Fujita, Marcos Silveira Buckeridge

University of São Paulo, Institute of Mathematics and Statistics, University of São Paulo, University of São Paulo, University of Illinois at Urbana-Champaign, Institute of Mathematics and Statistics, University of São Paulo, Institute of Biosciences, University of São Paulo

Abstract

The large amount of data now being produced by the omics techniques highlights a need of tools able to process and produce biological meaning from them. A holistic or systemic view to the object of study is a form of looking for emerging properties not present in its parts. The use of networks stands among the systems biology tools for analyzing relationships among variables and produces a more comprehensible view of a rather complex system. In Plant Sciences, network analysis have been used to probe transcriptomics, proteomics and metabolic data in a search for emergent properties of plant functioning. However, physiological approaches produce relatively low amounts of data and have been poorly integrated to omics data. Here we report the use of a relatively small data set on the physiological and metabolic behavior of the whole plant of sorghum in the form of networks. This afforded an emergent view of plant physiological behavior along a 24h period. We used data of photosynthesis parameters and sugar metabolites to construct correlation networks that afforded to view integrated physiological and metabolic transitions of leaves, stem and root of the plants simultaneously. The network topology analysis led to the visualization of the transitions of a higher diurnal connectivity related to photosynthesizing leaves and a central connected carbohydrate metabolism during the night, when organs establish mutual communication that is probably leading to growth. All data could be visualized in a single figure that show physiological behavior, without the need to analyze a large set of figures. Centralities (eigenvector, degree, betweeness and closeness) analysis pointed out the importance of starch as a central compound in the plant at certain times of the day. This type of analysis revealed deeper aspects of the system, such as detecting emerging centralities that only the whole plant displays, but the isolated organs do not. We are now starting to adapt an existing software (CoGA) - previously designed for comparisons of transcriptome patterns in humans - with the aim of producing a software capable to integrate plant physiological data (CoGA-Plant) which will afford fast and reliable evaluation of whole-plant physiological behavior. For that, other datasets related to trees, grasses and other plant species existing at LAFIECO (Laboratório de fisiologia ecológica de plantas) will be used. We also expect that this approach will be useful for other types of organisms, such as animals. This work is financed by FAPESP-Microsoft (2013/15571-3)

PATHChange: An R tool to identify differentially expressed pathways in Affymetrix microarray data

Carla ARS Fontoura, Enrico Giampieri, Gastone Castellani, José Carlos Mombach

UFSM, Universidade di Bologna, Universidade de Bologna, UFSM

Abstract

Statistical tools that deal with functional genomics and interactomics information (pathways) can provide insights on the causal origins of a perturbed state, for example, a pathological state and its normal (unperturbed) state. In terms of data analysis, deeper insights can be obtained from a combination of functional genomic with interactomic information. However, the use of a single statistical technique to infer significant changes in this type of analysis is not always satisfactory due to its inherent limitations, so a combination of two or more statistical methods is usually preferred. In this paper we present a PATHChange, a code written in R language based on the Bootstrap method, Fisher exact and Wilcoxon tests that improves the detection of differentially expressed pathways. PATHChange provides data normalization and statistical significance analysis of gene pathways alterations. As an illustration of the method, we applied the tool to a microarray dataset on colon pre-cancer and cancer from GEO database. The pathways used were obtained from the Ontocancro database that stores information on genome maintenance pathways that are commonly altered in cancers and age related diseases (`http://www.ontocancro.inf.ufsm.br`). Individually, statistical tests identify different sets of pathways with significant changes, however a small set of pathways (the intersection set) is identified as differentially expressed by all methods and has higher confidence of indicating a real change. We suggest this result is the best choice as an indicator of the effect of a perturbation. In addition, the intersection set yields an improved distinction of phenotype differences between colon pre-cancer and cancer in agreement with the literature.

A new, highly efficient strategy for decomposing population genetic structure and reducing consanguinity in non-random samples through a G-matrix based, centrality approach

Pablo Augusto de Souza Fonseca, Fernanda Caroline dos Santos, Mateus Henrique Gouveia, Thiago Peixoto Leal, Izinara da Cruz Rosse, Ricardo Vieira Ventura, Marco Antônio Machado, Marcos Vinícius Gualberto Barbosa da Silva, Maria Gabriela Campolina Diniz Peixoto, Eduardo Martin Tarazona-Santos, Maria Raquel Santos Carvalho

Universidade Federal de Minas Gerais, University of Guelph, Embrapa Gado de Leite

Abstract

Genome-wide association studies depend on samples composed by unrelated individuals. Therefore, strategies for reducing the relatedness level in the samples are crucial. The present study was carried out to investigate the impact of different resampling strategies on the genetic distance between individuals in a sample. 1036 Guzerá cows were genotyped using Illumina Bovine SNP50 and a subset of 11,264 markers from the whole set was used. Three resampling strategies were compared. For the first strategy, the pairwise Kinship coefficients (ϕ_{ij}) were estimated among all the individuals. All pairs of individuals showing a $\phi_{ij} \leq 0.1$ were eliminated (IBD strategy). After filtering, only 158 animals remained in this subsample. In the second strategy, the sample was modeled like a network, where each node was an individual connected to the others by edges, representing kinship coefficients >0.1. In the third strategy, the edges between the individuals were based on the values obtained by the genomic relationship matrix (G matrix) divided by 2. These strategies produced subsamples with 210 and 286 animals, respectively. These last two strategies allowed, by filtering against the higher centrality nodes, to eliminate all family clusters, without eliminating as many individuals as the IBD strategy. After these resampling strategies, only 21 cows were kept in all the three subsamples, suggesting that the differences among the subsamples obtained using these three different strategies depend not only of the number of individuals kept in the subsample, but also of who are the individuals kept. In order to measure the genetic distance among the animals in each subsample, the degree of similarity among the individuals was calculated based on a multidimensional scaling (MDS) method using the Euclidean distances between individuals based on the number of opposite homozygote genotypes. The MDS plot shows, for the whole sample and the three subsamples, that the animals in G matrix subsample maintained the larger number individuals with higher genetic distances. The G matrix approach uses allelic frequency to estimate the relationship coefficient. Otherwise, the Kinship coefficient only takes into account the number of alleles shared among individuals, resulting in a better identification of the most central and related individuals in the network. The results reported here suggest that the node selection algorithm based on the degree of centrality of a network is the better strategy for reducing relatedness. Moreover, the use of the Genomic matrix resulted in a better representation of the original sample.

Control Devices for Brain Waves

EDMAR ALVES COSME, Rosangela Silqueira Hickson Rios, TADEU HENRIQUE LIMA, MARLUCIA BEATRIZ LOPES PEREIRA

INFORIUM

Abstract

The human brain consists of nerve tissue and nerve cells which strategically act to conduct, to process, to interpret sensory stimuli and generate specific sensory responses when it is needed. For example, when a person wishes to turn on a light, it is generated an electrical stimulus which in milliseconds reaches the cerebral cortex located in the central nervous system. After the interpretation, an electric motor response goes back to the peripheral nervous system, going through neurons and motor nerves to trigger the muscles to contract and then, the person moves the arm to turn the light on. The electricity that propagates by the traffic of information on different types of waves, which depend on the different types of stimulus, and are captured by electrodes, as it happens in the electroencephalogram test, which evaluate the encephalon conditions through the registration of electrical waves. During stimulus processing, the area of the cortex responsible for this function concentrates a high rate of electrical activities due to the thousands of neurons mobilized to conduct, make synapses, decode and generate responses. The frontal lobe is the region responsible for the messages related to planning, initiative, reasoning and intuition, triggering an accumulation of electricity when a person wants something that needs to be planned, such as turning on the light, but the response generated only reaches the muscles if the person does not have physical limitations that prevent the stimulus path and the execution by the muscles. People with momentary or permanent physical disabilities and healthy cortex they process the sensorial stimulus, but do not implement motor responses because of the anatomical and physiological disability. In this case, triggering devices to turn on a TV, to change the position of the head of the bed or to ask for something, this represents a huge challenging. The possibility of capturing people's brain waves to enable devices to meet their needs, leads to the elimination of physical barriers and allows independence in daily activities. The work which was done, aimed to develop a mechanism to detect the brain wave frequencies and through that, enable the operation of any desired device. The prototype is based on a mechanism to uptake the waves related to the concentration and planning moments, it is composed by a transmission system (sensor, converter and Bluetooth transmitter) and a receiving system combined with hardware and software for system integration

Transcriptional Network Analysis applied to the 1-α-hydroxylase gene regulation in macrophages challenged by LPS

Romina Martinelli, Lucas Daurelio, Luis Esteban

Universidad Nacional de Rosario, Instituto de Biología Molecular y Celular de Rosario (IBR-CONICET)

Abstract

The vitamin D can be activated by the rate-limiting enzyme 1-α-hydroxylase (CYP27B1) synthesized mainly by renal cells. But other kind of cells can produce it and its regulation has different characteristics. In cells of the immune system, this enzyme is under the control of immune stimuli. For example, macrophage exposure to lipopolysacharide (LPS) has been implicated in the up-regulation of CYP27B1. The aim of this work is, by means of a system approach, to examine the complex interaction that involves the enzyme regulation in macrophages challenge with LPS. The transcription profiles of GSE40885 (Affymetrix Human Genome U133 Plus 2.0 Array) were downloaded from Gene Expression Omnibus (GEO). In this experiment, the RNA was isolated from alveolar macrophages of healthy men (n=14) by bronchoscope after the instillation of sterile saline or lipopolysacharide (LPS), and were divided into exposure group and control group accordingly. The raw data was preprocessed (correction, normalization) by RMA function of affy package of R 3.03 software in Bioconductor (`http://www.bioconductor.org/`). We applied Weighted Gene Co-expression Network Analysis (WGCNA) to detect clusters of highly co-expressed genes (modules). The weighted adjacency matrix assesses continuous connection strength ([0, 1]) according to β parameter for each group. Adhering to the scale-free topology criterion, α=6 was considered. The co-expression matrix and the topological overlap matrix (TOM) were constructed subsequently. With average linkage hierarchical clustering, gene modules were identified for each group. WGCNA analysis assesses the correlation between genes and considers degree of gene shared neighbors across the whole network. Moreover, WGCNA can provide connection strength between the genes, unlike general co-expression network. The module containing CYP27B1 was identified, and co-expression with c/EBPb and NfkB, both related with the 1-α-hydroxylase gene regulation, was found. Enrichment in Pathways using DAVID (`https://david.ncifcrf.gov/`) was done to get functionality information about this module. The gene list was enriched in 17 KEGG pathways, from which "p53 signaling pathway" and "Pancreatic cancer" were the most statistically significant. In addition, it was involved in the "Cell cycle" pathway. These biological activities are related to the well-known vitamin D anti-proliferative actions and result interesting to to find potential therapeutic targets.

Low cost portable electrocardiogram

MARLUCIA BEATRIZ LOPES PEREIRA, TADEU HENRIQUE LIMA, EDMAR ALVES COSME, Rosangela Silqueira Hickson Rios

INFORIUM

Abstract

The circulatory system has an impact on all cells, tissues and organs of the human body. The functions of oxygenating and nourishing have got a fundamental importance to life. The heart as blood ejection pump, needs to operate within an appropriate standards, complying with complex criteria for physiological stabilization. Medical exams seek to evaluate your situation and get possibility diagnosis, treatment and intrinsic care. The electrocardiogram is one of the most commonly used tests in cardiology, recording the generated electric waves, someone can judge the physiological condition of the patient. Despite of the importance, the exam shows limits of using, it costs about R150, 00*itisintra* − *patienthospital*, *incasethepatientisnotenrolledin*, *thepatientneedstogotoanelectiveconsultationandthensetatimefortheexamaccordi* it has got the size of 15cm3, it can be used as monitoring home , tours and journeys, easy to handle, it does not require a trained professional to operate it; It can be triggered when the patient feels the need; it enables to generate results that can be routed and monitored by teleprocessing real-time signals. The developed prototype used the microcontroller Arduino NANO and showed positive results, indicating the perfect working order. To develop this system was created a hardware associated with a software, hardware, compound by microcontrollers allow the communication with the electrodes placed on the patient to generate the data. The software will do all the processing and plotting a graph in real time. Thus, it aims to encourage the dissemination and implementation of this technology in the market, increasing the early diagnosis and treatment, as well as reducing the severity of cases and the number of deaths from heart disease.

SigNetSim: an e-Science framework to design and analyse dynamical models of molecular signaling networks

Vincent Noel, Marcelo S. Reis, Matheus H.S. Dias, Layra L. Albuquerque, Fabio Nakano, Junior Barrera, Hugo A. Armelin

Instituto Butantan, União Educacional do Norte UNINORTE, Universidade de Sao Paulo

Abstract

One of the current major challenges in Molecular Cell Biology is to properly analyze the big data yielded by modern high-throughput omics techniques (e.g. genomics, transcriptomics, proteomics). Classical biological intuition and qualitative, static interactome schemes are no longer sufficient to study underlying dynamical biological mechanisms from such huge amount of data. On the other hand, computational approaches are suitable to tackle this and other timely goals of nowadays biological research. To this end, e-Science is an emergent discipline within Computational Sciences focusing on developing theories and tools to allow biological scientific investigation on a computationally intensive environment. Therefore, the objective of this work is the development of SigNetSim, an e-Science framework to assist mathematical modeling and computational analysis of molecular signaling network kinetics. This framework will allow the usage of big data and also the traditional low-throughput omics data (e.g. Western blot experiments) into modeling and validation processes. It uses the standard representation for systems biology dynamical models (SBML), enabling the utilisation of the numerous existing models and the portability of produced models to other compatible softwares. Users can access the framework locally, through command-line or graphical interface, or remotely through a web interface. As a case study, we modeled the Ras/MAPK signaling pathway in mouse Y1 adrenal tumor cell line and showed that [K-Ras-GTP] relatively high steady basal levels, a condition experimentally observed in Y1 cells, are achieved only with the inclusion in the model of SOS and also an additional guanine exchange factor (GEF). To experimentally validate this hypothesis, we probed Y1 cells for expression of additional GEFs by RT-PCR, confirming the expression of two additional GEFs. The SigNetSim e-Science framework was successfully used to model different sets of experiments. Presently, we are working on the design of a kinetic model that explains the crosstalk between the Ras/MAPK and PI3K/Akt signaling pathways in the same cell line. In the mid-term, we intend to use high-throughput quantitative proteomic data in this modeling task, improve the framework to perform assessment of different hypotheses for the model structure, and add more analyses methods. Supported by CNPq and by grants #12/20186-9, #13/07467-1, and #13/24212-7, São Paulo Research Foundation (FAPESP).

Genome-scale metabolic network reconstruction of the bacteria Burkholderia sacchari

Paulo Alexandrino, Luiziana Silva, José Gomez, André Fujita

University of São Paulo

Abstract

Burkholderia sacchari is a gram-negative bacteria that was isolated from a sugarcane plantation in Brazil. The bacteria has drawn attention both because of its ability to accumulate high yields of polyhydroxyalkanoates (PHAs) as storage material and also due to its versatility in using carbon sources comprising glucose, sucrose and xylose. PHAs are biodegradable polyesters notable for its potential to substitute some petrochemical plastics, thus generating a positive environmental impact. However, PHA production is still not an economically feasible process and more advances in PHAs fermentation are needed. In this context, the field of metabolic engineering points out as a means to manipulate genetics with the aim of producing chemicals, where genome-scale metabolic networks are used to simulate an organism's metabolism, serving as a blueprint for genetic engineering. The aim of this work was to generate a genome-scale metabolic network reconstruction of Burkholderia sacchari. To reach this objective, we have sequenced the bacterial genome using 454/Roche technology and the obtained reads were preprocessed using Prinseq and assembled with Newbler. The set of contigs was then submitted to RAST, an automatic annotation pipeline, and the resulting annotated genome was used as input in the ModelSEED pipeline, which generated the initial metabolic network reconstruction. Manual curation of this first draft was carried out in two steps. First, MetaNetX was used to ensure a controlled vocabulary for the network's components. Then, the reconstruction was represented in a graph database, where all the network entities could be compared to other species reconstructions and manually edited. The result of the present work is the first genome-scale metabolic network reconstruction of Burkholderia sacchari.

Influence of the Cell Volume on the Dynamic of the Mammalian Cell Cycle

Alessandra Cristina Gomes Magno, Itamar Leite de Oliveira

Universidade Federal de Juiz Fora

Abstract

The goal of this work is to perform numerical simulation of the mammalian cell cycle model taking into account the cell volume. The cell division occurs according to the work of Chen et al (2000), that is the moment in which the cyclinB/Cdk1 species is totally degraded and it reaches its minimum value. In this time, the cell volume drops, in average, to its half; simulating the cell division into two daughter cells with equal cytoplasmic (cytoplasmic organelles) content. Gérard and Goldbeter (2011) proposed the equations used in the mathematical model and they are valid only when the volume, where the reactions occur, is constant. When the cell volume is not constant – that is, it varies along to the time according to a differential equation –, that original equations are not valid anymore. Therefore, every equations were modified from the mass conservation principle and they considered a volume that changes with time. Through this approach, the cell volume affects all model variables. It modulates the synthesis of the cyclinD/Cdk4-6, cyclinA/Cdk2, cyclinB/Cdk1, cyclinE/Cdk2 complexes and it affects the E2F transcription factor and the Cdc20 protein. Two different dynamic simulation methods were accomplished: deterministic and stochastic. In the stochastic simulation, the volume affects every model's parameters which has molar unit, whereas in the deterministic one, it is incorporated into the differential equations. In deterministic simulation, the biochemical species may be in concentration units, while in stochastic simulation such species must be converted to number of molecules that is directly proportional to the cell volume. In order to find the approximated solution of the deterministic model, the fourth order Runge Kutta method was coded in the C programming language. As for the stochastic model, the Gillespie' Direct Method with variable volume by Kampen and Godfried (1992) was implemented. In both simulations, the obtained results were according to that original ones.

Integration of gene expression data of Leishmania infantum via biological interaction networks

Frederico Guimarães, Leilane Gonçalves, Juvana Andrade, Daniela Resende, Pascale Pescher, Gerald Späth, Silvane Murta, Douglas Pires, Jeronimo Ruiz

Centro de Pesquisas René Rachou – Fiocruz Minas, Instituto Oswaldo Cruz, Institut Pasteur

Abstract

Leishmania parasites cause a broad spectrum of clinical diseases known as leishmaniasis. Annually, approximately 1.3 million new cases are reported and many therapeutic failures occurring due to development of drug resistance have been observed. We have performed a comparative transcriptomics analysis of L. (L.) infantum chagasi line (MHOM/BR/74/PP75), considering sensitive and resistant strains to potassium antimonyl tartrate (SbIII), using NGS Illumina Sequencing. In order to investigate the differential gene expression associated with drug-induced stress response and SbIII-resistance mechanisms, we have compared SbIII-treated and non-treated samples of each strain. Additionally, a protein-protein interaction (PPI) network was created and the genomic positions of proteins translated from these differentially expressed (DE) genes in Leishmania infantum were identified for further investigation. In the initial analysis process, TopHat 2 was used for mapping the reads against the reference genome and DESeq 2 was used for DE statistical analyses. When comparing the data from SbIII-resistant strain against SbIII-sensitive strain (non-treated, used as control), the analytical pipeline allowed the identification of 200 differentially expressed genes. For PPI analysis, protein network data for Leishmania infantum JPCM5 was downloaded from the String database (version 10). A parsing script for filtering scores greater than 900, on categories ”experimental” or ”database” was developed. An interaction network was then created, using the Cytoscape software. The network was further annotated with data derived of DE genes, obtained earlier, with an adjusted p-value smaller than 0.05 and log2 fold changed greater than 1.2. We then focus our attention on searching for protein groups in the network (indication of co-regulation) as well as ”hub” proteins. In a later stage, metabolic pathways associated with the proteins translated from differentially expressed gene, mapped on network, were annotated using KEGG and functional enrichment analysis (biological process, molecular function and cellular component), using the BinGO Cytoscape plugin. We have carried on further analyses on network structure, using Cytoscape internal tools. Amongst the identified protein groups and hubs we have found several proteins related to phospholipid glycerol metabolism, aminobenzoate degradation and biosynthesis of fatty acids.

ATiNEU, a proposal for a general a general purpose on-line tool to manage digital brain atlas

Lucas Felipe da Silva, José E. O. da Costa, Anderson Souza, Wilfredo Blanco

State University of Rio Grande do Norte (UERN)

Abstract

A Digital Brain Atlas (DBA) can be constituted by sequences of digital images of one or more representations of the brain. DBAs play a very important role to explore the morphology of brain regions. Recently, the DBAs incorporate an important feature, they can be seen as containers, a computational framework that integrate several data sources (images, genetic dada, functionality and time dynamic among others) to describe and understand better the central nervous system of species. Although there are very well know international projects, we did not find any of them developed in Latin America, including Brazil. This article presents the methodology for the creation of ATiNEU, an on-line framework system to storage, manipulate, visualize and annotate DBAs images. The system combines the advantages of highly organized relational database and a front-end Web application. The Database was created and managed by MySQL. The Web application interface was implemented based on Model–View–Controller (MVC) software architectural and using HTML, CSS, PhP and JavaScript. Images were stored outside of the database and to manage their visualization efficiently, different resolution copies were created. Through the annotation tool, images can be graphically annotated (text, points, lines and regions) and this feature was developed using the canvas element from HTML5. The relevance of this work is sustained by the fact that facilitate the collection and sharing on-line heterogeneous data from various central nervous systems. The graphical annotation tool also provides biologists and neuroscientist with a flexible instrument to perform diverse annotations. In this context, we believe that this project will be pioneering in Latin America, not only for consolidate data from nervous system, but and not less important, the creation of a basic computational infrastructure for handling data from the Digital Brain Atlases.

GenSeed-HMM: a tool for progressive assembly using profile HMMs as seeds - application in virus discovery of Alpavirinae from metagenomic data

João Marcelo Pereira Alves, André Luiz Oliveira, Tatiana Orli Milkewitz Sandberg, Jaime Moreno-Gallego, Liliane Santana Oliveira, Alan Mitchell Durham, Paolo Marinho Andrade Zanotto, Alejandro Reyes, Arthur Gruber

University of São Paulo, Universidad de los Andes

Abstract

In a previous work, our group introduced GenSeed (Sobreira T.J. & Gruber, A. - Bioinformatics 24: 1676-1680, 2008), a program that implements a seed-driven progressive assembly, a robust and reliable approach to reconstruct specific sequences from unassembled data, starting from short nucleotide or protein seed sequences. GenSeed-HMM is a completely revised and extended version of GenSeed: in addition to nucleotide and protein sequences, profile HMMs can now be used as seeds for sequence reconstruction. The program can use any one of a number of sequence assemblers, namely CAP3, Newbler, AbySS, SOAPdenovo, and Velvet. As a proof-of-concept and to demonstrate the power of HMM-driven progressive assemblies, GenSeed-HMM was applied to previously published metagenomic datasets in order to search for ssDNA bacteriophages from the Alpavirinae subfamily (Microviridae family). A dataset of Microviridae proteins was used to perform multiple sequence alignments of VP1 and VP4 proteins and regions specific to Alpavirinae were used to build profile HMMs using hmmbuild. These HMMs were then used by GenSeed-HMM (running Newbler) as seeds to reconstruct viral genomes from sequencing datasets of human fecal samples (SRA codes SRX028823 to SRX028827). All contigs obtained were automatically annotated with the EGene2 platform (Durham, A. et al. - Bioinformatics 21: 2812-2813, 2005) and taxonomically classified using similarity searches as well as phylogenetic analyses. The most specific seed, VP1R4, enabled the reconstruction of 45 genome sequences from the Alpavirinae subfamily. A comparison with conventional (global) assembly of the same original dataset using Newbler revealed that GenSeed-HMM outperformed global genomic assembly in the several metrics employed. Because assembly is performed in multiple steps and relatively few reads are used on each cycle, the process demands low computational resources in terms of processing power and memory allocation. GenSeed-HMM provides a fast and simple implementation to run progressive assembly pipelines that are directly targeted at sequences of interest, generating better assemblies and potentially accelerating the pace of novel virus discovery. Other potential applications include the specific assembly of episomal elements, gene family/diversity studies, and detection of antibiotic resistance genes in metagenomic data and their context (plasmid or chromosomal).

Evaluation of the InterProScan interface from the perspective of systems information

Rafael Moreno Ribeiro do Nascimento, Maurílio José Inácio, Laila Alves Nahum

Faculdade Infórium de Tecnologia, Universidade Estadual de Montes Claros,Faculdade de Ciência e Tecnologia de Montes Claros, Centro de Pesquisas René Rachou (CPqRR)

Abstract

Abstract Background: Information technology is key for the integration of other areas of knowledge. Besides, computers are being used intensively for storing and organizing protein databases. It would be certainly very difficult to work with the current volume of information and data if computer resources were not used. Information technology is also an important means for the user to get access to resources of these databases. This study aimed to analyze and propose a new interface to a system used for the study of proteins, InterProScan, from the perspective of an information systems professional. The assessment was done using questionnaires, one addressing the aspect of interaction with the user and the other on the organization of information in the interface. Results: As a result, it was realized that the system has functional usability features that comply with good usability practices, however, it has some deficiencies as to how the information is displayed. In addition, a prototype as a suggestion of a new interface has been developed for the software as well as suggestions of features that are not implemented in this study. The prototype suggests new features for InterProScan as the visualization of functional domains using a bar graph so that the user can compare the percentage of coverage of each domain in relation of the original protein; a more direct way has been implemented so that the user has access to information using a simpler interface, and may also choose whether to use the signature or the particular protein identifier as input. Through the suggested implementation, the system provides a better user experience and saves on computing resources. Conclusions: How to prospect other related works could be used to extract knowledge from information returned by InterProScan by using computational intelligence techniques such as neural networks. This type of resource is already recognized for this purpose, being an efficient technique in the text mining problems in several areas. We conclude that the results of this study may contribute to possible improvements of InterProScan system or serve as a basis for future research related to improvements of InterProScan system or other computer systems applied to biology.

An distributed environment for data storage and processing in support for bioinformatics analysis

Leandro Cintra

Embrapa - Brazilian Agricultural Research Corporation

Abstract

Nowadays, new technologies on data generation are amplifying in an unthought manner the amount of biological information accessible for analysing. This present us scenarios at which storage spaces and processing capabilities are bottlenecks on the computational system. The issue related with processing can be addressed with a computer cluster in which is possible to execute the tasks of an analysis in parallel. However, the storage issue is not well addressed for huge amounts of information with traditional tools. Normally, the clusters systems use NFS (network file system) to provide an unique information repository in the computation environment and this will have some disadvantage: a) storage space is limited by the capability of the server, which means that it will not be sufficient for those cases with great storage demand b) data throughput is limited by the server capability c) and all the system will be nonoperating if the file server go down by some reason. In this work, we investigate the use of a distributed file system (DFS) for data storage; associated with a distributed resource management (DRM) for control the parallel tasks execution. With a DFS system the environment can scale for petabytes of storage and operate in high throughput. Our environment was configured with six machines each one with 4xIntel Xeon E5-4620 (32 cores), 512Gb Ram and 887Gb of usable storage space in RAID 6. They were connected with a Gigabit ethernet network. These nodes with low storage capability were used as testing and specialized storage nodes are being provided with about 40Tb each one. We used the GlusterFS system as DFS and the Gridengine system as RDM. Bioinformatics tools with intensive IO activities were used in benchmarks of the system. Among them are blast, interproscan, SAM/BAM tools and the genome assembler MaSuRCa. Our tests were made considering local file system, NFS file system and GlusterFS file systems. The results indicated that the distributed storage system is stable, resilient and has potential to be used in production environment. Parallel environments based on processing of distributed tasks with distributed file systems are a promise approach for bioinformatic's demands and would be considered in projects working with big amounts of biological data.

ONE STEP TO UNRAVEL AND DESIGN PRIMERS FOR CONSERVED MICROSATELLITES IN SEVERAL GENOMES

Marcelo Soares Souza, Lucas Soares de Brito, Alexandre Alonso Alves, Eduardo Fernandes Formighieri

Brazilian Agricultural Research Corporation,

Abstract

Despite the predominance of high throughput technologies, microsatellites are still important, for example, as molecular markers. Typically, people designs primers to target specie and then test amplification in other species hoping to get lucky, usually wasting time and resources. Our goal is to develop a pipeline that automates the entire process, from microsatellites detection to the primer design in conserved regions between different genomes. To this end, we are using recognized tools such as Tandem Repeats Finder (TRF), NCBI Blast+, Clustal and Primer3 for the different stages of analysis. We are using Python/Django for programming pipeline and web interface, and PostgreSQL as SGDB. The inputs are one or more genomes (or assembles) in fasta format and the parameters definition (we will provide default values). The first genome (G1) is the target, in which primers will be designed. We expect other genomes (G2, G3 etc) been provided in order of taxonomic proximity to the G1, being G2 the closest and so on. Any of the genomes can be set as reference, to assessing distance between SSRs. The tool, still unnamed (you can suggest/vote for a name in the poster session), will first run TRF and TRAP for all genomes. The results (including microsatellites (SSRs)) will be filtered (e.g. by class – dinucleotides, trinucleotides etc), and selected information for each SSR selected will be loaded at the database (DB). Each filtered G1 SSR, including upstream and downstream borders will be compared (Blast) to the other genomes, and the similar regions will be scanned (SQL search) to find similar SSRs. When found, the sequences of the regions will be aligned (Clustal) to find conserved regions that will be considered preferential regions at primer pair design (Primer3). According to previous steps, the tool will attempt to find primers in regions conserved in as many genomes as possible, decreasing the number of genomes when Primer3 cannot find good primers. In all stages, selected information will be loaded at DB. At least, an advanced search (web interface, with access control) will be developed to deliver primer pairs report, including several search terms, such as: number and class of SSR, amplicon length, melting temperature, genomes considered at primers design; and minimum distance between each two SSRs. After development and testing, we intend to submit the full work as a technical or a research paper, and make available the source code.

THE NEEDLEMAN-WUNCH PYTHON SCRIPT

Rodrigo Langowski, Marthin Borba, Alessandro Brawerman

UFPR

Abstract

The Needleman-Wunsch algorithm is commonly used in Bioinformatics to align two nucleotides or amino acids sequences of similar sizes. The objective is to find the best possible alignment between those two sequences. This method is referred as a global alignment that runs the full length of the two sequences in order to get the best result. An implementation of this algorithm was made using the Python programming language, which is a very simple and easy language for beginners and yet is at the top 5 most used programming language, according to the TIOBE classification. The program automates the comparison of the two amino acid sequences and presents, with success, the best possible alignment together with their identity. With Python, we can easily transform this program in a library module and share with developers around the world so they can import and effectively use our implementation with any sequences they need, independent of their size, without the need of redoing the code. Our Python program requires some input from the user, as such, setting the gap (penalty for a given insertion or deletion), calculating match/mismatch weights and informing the two sequences to be analyzed. After performing the calculations, comparisons and analysis, the algorithm returns a table containing a parcial result, with the calculated values for match/mismatch, along with the gap penalties set previously. Following these first step, the algorithm performs a traceback in the table in order to find the best possible alignment. At the end, the alignment sequence, together with the identity, are presented. The script was designed to optimize the global alignment process between two sequences, facilitating the matrix calculation and the alignment achievement, making manual or logic effort unnecessary, which would also be impractical when dealing with long sequences of amino acids. Supported by: CAPES.

TOWARDS THE DEVELOPMENT AND VALIDATION OF HIGHLY EFFICIENT PIPELINE TO PERFORM INTREGRATED ANALYSIS OF REPETITIVE REGIONS IN COMPLEX GENOMES

Lucas Soares de Brito, Jaire Alves Ferreira Filho, Marcelo Soares Souza, Manoel Teixeira Souza Júnior, Alexandre Alonso Alves, Eduardo Fernandes Formighieri

Brazilian Agricultural Research Corporation

Abstract

Plant genomes are highly populated by repetitive sequences, which may represent, more than half of the whole sequence. In fact, 50-70% of the plant's genome sequence are in fact composed of the most common such as transposable elements and mini/microsatellites. The first are usually mutagens, and may be used in genetic engineering while the last are usually converted into molecular marker to aid breeding programs in a variety of ways. Despite, its possible applications in breeding programs, these types of repetitive sequences often require different tools to discovered and analyzed, and a complete genome survey for such elements is quite laborious. Therefore, we are working in a tool/pipeline to integrate different software solutions used to find and classify all types of RR. Our aim is to reduce the complexity of the analysis itself, making such task less consuming in terms of time and human efforts. The pipeline is being developed with Perl language, and includes well-known software like Tandem Repeat Finder (TRF), RepeatModeler, RepeatMasker and NCBI Blast+. The initial inputs are genome/assemble sequences (FASTA format); and a configuration file, which specifies the parameters for each individual tool to be used (default values are being defined). All software and languages must be previously installed on Linux environment. Briefly the main components/steps of our pipeline, and the respective task performed by it are: (i) TRF/tandem repeats analysis; (ii) TRAP/summarize TRF results; (iii) LTR Finder/Long Terminal Repeats analysis; (iv) RepeatModeler/repeat boundaries and family relationships analysis; (v) parsing and concatenation of previous results (scripts Perl); (vi) creation of local databases (makeblastdb); (vii) NCBI Blast+/repeats' classification (against local and external DBs); (viii) parse of blast tabular results (e.g. evalue); and (ix) RepeatMasker/annotation. Steps (i) to (viii) were implemented and underwent initial tests with a local assemble of an Elaeis oleifera (oil palm) genome draft. Now we are working at the RepeatMasker analysis, performing more tests, to improve speed and presentation of the results. Finally, we intend to make the source code available to all bioinformatics/genetics community and publish a brief description of such integrated tool, and its highlights.

PFSTATS: A GUI-based software for protein family analysis by conservation detection and decomposition of residue coevolution networks

Néli José Fonseca Júnior, Lucas Bleicher, Afonso M.Q.L.

Universidade Federal de Minas Gerais

Abstract

Structural and functional insights about protein families can be trivially obtained by conservation analysis. Additional and potentially useful information can also be achieved by correlation analysis. Our group has recently proposed a method to obtain functional sub-class determinants in protein families, called Decomposition of Residue Coevolution Networks (DRCN). DRCN is a sequence based method for analysis of protein families represented by multiple sequence alignments. We present a GUI-based software for protein family analysis using DRCN and conservation analysis. The algorithms were grouped in order to have a robust and intuitive application to analysis of homologous proteins. The software has a user interface developed with the QT framework and consists of five main modules. The first one is the alignment filtering, which is applied in three steps in order to remove possible fragments, poorly aligned sequences and phylogenetic bias. The final alignment is used for further statistical analysis, starting by conservation, which includes output to a B-factor modified PDB (changed to conservation values) if a structure is available. The next module uses a bootstrap-like method to estimate the smallest sub-alignment size for correlation calculation, which also considers cutoffs for minimum correlation scores and target frequencies. The result is a table of correlated and anti-correlated pairs followed by their correlation scores, which can be understood as a correlation network where vertices are residue-position combinations and edges are present if there is a statistically significant correlation or anti-correlation between them. This network is then decomposed into communities – sets of residues that are highly correlated among themselves, but not to the rest of the network. Finally, PFSTATS has a module that search for information described in the UniProt database for all conserved and correlated residues on all specified proteins. It also allows the visualization of the results in the user interface itself, as well as export to TXT, CSV, XML, HTML file formats.

A hybrid architecture for databases in bioinformatics workflow

Iasmini Virgínia Lima, Maristela Holanda, Maria Emilia Walter

University of Brasilia

Abstract

Experiments in bioinformatics are usually developed as scientific workflows comprised of different programs that generate a large volume of data, from millions of genomic sequences generated by high-throughput sequencers. Scientific workflows are scientific computational experiments composed of several programs, each with specific input data and parameters, combined to solve a problem. Bioinformaticians can execute the same workflow many times, with different parameters and databases, to compare the obtained results, and refine the analyses. Within many storage architectures for biological information presented in the literature, there are those based on systems managed by relational databases, and more recently, systems based on non-relational data bases, also known as NoSQL (Not only SQL). NoSQL databases are classified in different models, e.g.: databases directed to documents, key/value, columns and graphs. In this context, we propose to develop a hybrid architecture for databases, with relational and non-relational (NoSQL) database system managers, to store data originating from the execution of a bioinformatics workflow, seeking to maximize performance by choosing the best storage for each kind of information generated in each phase of the workflow. Important features are high scalability, distributed processing, ability to handle both structured and unstructured data, availability of NoSQL database, and some properties of relational databases, e.g., consistency. Here, we deal with provenance data, i.e., both information generated by the execution of the workflow, and data generated at each phase of each experiment. We intend to analyze the performance of the proposed hybrid architecture, regarding both to the volume of stored data, and the efficiency of an execution of the workflow.

Improving automation, reproducibility and installation of genomic analysis pipelines with Docker

Marcel Caraciolo, Filiphe Vilar Figuereido, Victor Monteiro

Genomika Diagnósticos

Abstract

Bioinformatics pipelines usually rely on a combination of several components and deploying them incurs substantial configuration and maintenance burden. Genomics and variant analysis pipeline is normally difficult to install, configure and deploy. Moreover in cluster environments the software must be deployed in several nodes (machines), which may raise inconsistent and serious reproducibility issues. At Genomika Diagnósticos laboratory, we tackled this issue with a scalable and repeatable approach using Docker containers (lightweight virtualization). This approach provides several advantages: efficiency, portability, versioning and reproducibility. Encapsulating NGS workflows working in containers, a user can quickly deploy any pipeline version in any environment (e. g. operating systems, workstations, clusters, clouds) and overcomes several issues from common used approaches with virtual machines (VM's). VMs lack portability, have substantial overhead (disk, CPU, RAM) and require allocated resources to be provisioned statically. Docker is a open-source software, it isolates the tools and software involved in processing, and makes easier to recreate a snapshot of the current environment of the pipeline for reproducibility without manual re-installation of specific versions of software. Our goal with this poster is to share our best practices and experiences for developing, distributing and running pipelines encapsulated in containers using Docker. We will give a brief introduction to the main concepts in the Docker programming environment and how variant analysis pipelines can be used together in order to deploy and run pipelines on multiple platforms in a repeatable manner. Additionally, we will show our current framework architecture that we are developing to improve our ability to build automatically versioned pipeline containers.

SEMI-SUPERVISED MACHINE LEARNING APPLIED TO MEDICAL DIAGNOSTICS

Diego Henrique Negretto, Erik Aceiro Antonio, Maurício Bacci, Milene Ferro, Fabrício Aparecido Breve

Universidade Estadual Paulista Júlio de Mesquita Filho, Universidade Federal de São Carlos

Abstract

Data mining applied to Medicine is an emerging and important field to detect new prognostics and to understand the classification of a particular disease. Thus, the main concept of medical technology is to learn about the disease characteristics and to provide diagnosis for patients in uncertain future disease phases, therefore Machine Learning (ML) play an important role to assist the analysis and evaluation of the raw dataset. In this context, Semi-Supervised Learning is a ML technique that provides mechanism to training dataset — labeled and not labeled data — which can minimize the time spent for labeling process aiming an analysis for dataset more feasibility for human expert. This work involves graph-based semi-supervised learning algorithms such as Particle Competition and Cooperation (PCC), Label Propagation (LP), Label Neighborhood Propagation (LNP), and Local and Global Consistency (GLC) which were developed on MATLAB. To evaluate accuracy of the classification three different datasets were used — Breast Cancer Wisconsin (BCW), Statlog (Heart), and Parkinsons, which are available in the UCI repository (Machine Learning Repository). These datasets were grouped into five different samples of labeled data (3%, 5%, 10%, 15%, and 20%). For each sample, it were selected ten subsets, ensuring that at least one example of each class were present. The PCC algorithm presented best results (correct answers) for all samples in BCW (97% to 97.99%) and Heart (64.7% to 73.8%) datasets, as well as 3% (80.15%), and 20% (87.4%) for the Parkinsons dataset. On the other hand, the GLC algorithm showed the best result for the 5% (82.8 %) and the LP algorithm was better for 10% and 15% (85.1% and 85.9% accuracy) for these dataset. In summary, we showed that the semi-supervised algorithms were able to achieve good accuracy — in comparison with other traditional literature results — when applied in the medical datasets used and this can be an interesting research area. Therefore, this result suggests that using Semi-Supervised learning we can minimize the gap existing on human expert analysis and machine expensive dataset used in data mining and data analysis.

Storage and recovery of dairy cattle genotype data from the data science approach

Rennan Silva, Fernanda Almeida, Wagner Arbex

Federal University of Juiz de Fora, Embrapa Dairy Cattle

Abstract

The genotyping process consists of the identification of molecular markers, which may vary from individual to individual. Usually, the records that identify these markers are stored in big text files and contains several informations about each individual, like the animal number and the values associated for each marker. There are different patterns to present the genotyping results, depending on the platform chosen for the job. Usually, the data gathered from this process are provided in text files. Analytical tools are not used to treat this kind of data because they are in a very large volume. There are several limitations on relational databases when data is big and unstructured. The goal of this abstract is to propose a way to store the results of the genotyping process (cattle SNP) and allow this information to be accessed from a middleware of ontologies. Besides, this database should provide a background to cross information about animals, individuals, SNPs, samples, etc. This model foresee the creation of an ontology to map some characteristics of the domain and simplify the queries on a genotype database. This ontology will be used on every possible slice of this process, since the data collect to the queries, enabling the data science on this process. Furthermore, this database will store the description of the processes that the data passed by until the moment they are stored, allowing their provenance. Once the proposed database does not fit on the basic principles of relational databases (atomicity, consistency, isolation and durability), it is necessary to implement this work in a less conventional feature, like NoSQL. This database will not need the delete and update procedures, so it may be off the normalization standards to benefit performance in queries. This work will simplify the access to genotype information and may help future works that depends on research over a molecular marker database in bovine dairy cattle.

DATA VISUALIZATION FOR SEQUENCE COMPARISON

DCB Mariano, TS Correia, JRPM Barroso, RC de Melo-Minardi

LBS: Laboratory of Bioinformatics and Systems - Federal University of Minas Gerais.

Abstract

Background: Biological data visualization provides ways organize, visualize and to analyze large collections of data on different perspectives, and thus, obtaining comprehensive information. Important use of data visualization in Bioinformatics are: comparisons among protein structures, analysis of transcription variation, comparisons among partial sequences or complete genomes (synteny between genomes). Synteny consists in genetic blocks which are conserved in order within two sets of chromosomes of different organisms, when compared with each other. Comparisons among sequences is one of the most common tasks on Bioinformatics, however the majority of the tools for this purpose show the results in textual formats. We believe that meaningful visualizations can improve the analysis of synteny comparison among genomes. Results: In this work, we propose a tool for comparison among sequences using Biopython library and other Python graphic library (Python Imaging Library - PIL). We also present visualizations for comparison between complete small genomes, that perform detections of genes position, syntenic regions, genomic inversions, and also includes markings in repetitive regions, such as regions that codify ribosomal RNA. Our tool realize comparisons using the software BLAST and can be executed at command line interface. It has as input GenBank files of different organisms, and as output, vector images in PDF format. In the synteny visualization result, we can detect synteny between genomes through segments of straight which interconnect the rectangles that represent the genomes. We use colors to distinguish different types of biological structures. Conclusion: Meaningful visualizations can help to discover important knowledge about biological process. Our tool provides simple visualizations that help to analyze differences among sequences. We believe that the proposed visualizations will be very useful for bioinformaticians. The is available at `http://thames.dcc.ufmg.br/bioview`.

POTTER: A WEB TOOL FOR PROTEIN POINT MUTATION MODELLING AND ANALYSIS

JRPM BARROSO, DCB Mariano, TS Correia, Rodrigues L, A Fassio, P Martins, C Leite, TJ Sousa, F Póvoa, RS Ferreira, L Bleicher, RC de Melo-Minardi

LBS - Laboratory of Bioinformatics and Systems. Department of Computer Sciences. Federal University of Minas Gerais

Abstract

Background: Point mutations, or single base modification, are mutations that affect one position in genes and may cause changes in protein structure conformation, and can affect its stability and function. Recently, several works have related point mutations to diseases, inducing changes in enzymes' efficiency and, in some cases, loss of function or inactivation. Thus, in silico simulation of mutations is important to predict effects in the structure and function of proteins. Although command line applications run generally fast, many are the obstacles faced by those who will try to use them in their research, such as the variety of operational systems used by the users that distinguishes platform or version, or often the lack of libraries and dependencies required to install the tools. Results: In this work, we propose a Web tool, called POTTER (PrOtein muTaTion viewER), for point mutations analysis and visualization in wild and mutant proteins. POTTER was developed using Python, PHP and JavaScript, and it is implemented in three different modules: (i) mutation impact modeling using the Modeller software; (ii) three-dimensional structure visualization using the 3DMol.js framework; and (iii) contact analysis using in-house scripts developed with Biopython. Input of the application is a PDB file (Protein Data Bank), the new residue used in the mutation, and the residue position in the sequence. The result of mutation can be viewed in parallel with original protein enabling a more detailed analysis of the mutation and the impact caused. The tool also enables the download of the generated PDB file. Conclusion: The friendly interface of POTTER allows an easy modeling of the mutations in proteins as well as a adequate to visualization of the results. A bioinformatician without a great specific knowledge of hardware and software can perform analysis of mutations easily. We plan in the future to implement molecular dynamics simulations and the respective result analysis to improve the mutation impact analysis. The application is available at `http://thames.dcc.ufmg.br/potter`.

DETECTING BETA-GLUCOSIDASES WITH HIGH CATALYTIC EFFICIENCY FOR CELLULOSE DEGRADATION USING SINGULAR VALUE DECOMPOSITION

TS Correia, JRPM Barroso, DCB Mariano, RC de Melo-Minardi

UFMG

Abstract

Background: Beta-glucosidase (BGL) is an important enzyme for the production process of lignocellulosic biofuels. It is responsible for cellulose degradation in the glucose used in the fermentation process. BGL has been classified into several subfamilies of glycosyl hydrolases, such as GH1, GH3, GH5, GH9, GH30, and GH116. Studies have showed members of the GH1 family are more efficient for cellulose degradation. Despite traditional methods, such as BLAST and HMM, have been efficient to classify sequences family, this classification is not totally sufficient to determine the enzyme efficiency catalytic. In this work, we propose the use of Singular Value Decomposition (SVD) to detect BGL with high catalytic efficiency. SVD is an useful technique of linear algebra with a large number of practical applications in information retrieval and detection of non-obvious relationships among similar elements. For evaluate our method, we extracted 4,932 sequences from UniProtKB database in FASTA file, 217 sequences from SwissProt used as gold-pattern, and its respective families in a tabular file. We analyzed k-mer information, calculated SVD and extracted the three main columns with singular values. These values were used to plot a tridimensional graph, where each point represent a protein. Then we calculated contacts among points using delaunay algorithm. The contacts were used to predict the family of the proteins obtained from UniprotKB based on our gold-pattern, and then compared with the values in the tabular file. Results: We detected that our method hit 98% of the family predictions. We hypothesized proteins from UniprotKB represented by points in tridimensional graph which present contacts with proteins from SwissProt with high efficiency for cellulose degradation validated in bench are strong candidates to also present a high efficiency. For example, the protein ”1,4-beta-D-glucan glucohydrolase” from Thermotoga neapolitana KCCM41025 is presented in literature as an enzyme with a high catalytic activity. Our method detected 2,580 proteins unreviewed clustered near to ”1,4-beta-D-glucan glucohydrolase”. These proteins are potential targets to enzymatic kinetics experiments. Conclusion: In Bioinformatics several works have presented the SVD use as an efficient approach for proteins clustering. In this study, SVD can be used successfully to detect proteins family of beta-glucosidases, and we also proposed a new use for detect enzyme efficiency. We developed a class in Python to facilitate the running of SVD for biological data. The scripts and case study data are available at `http://thames.dcc.ufmg.br/biosvd`.

Comparison of The Main Tools to Identify Inconsistencies, Manipulation and Research Files in Protein Data Bank

Wellisson Gonçalves, Raquel C. de Melo-Minardi

UFMG

Abstract

We compare the main tools for treating biomolecules files found in the Protein Data Bank, building a critical analysis of its features search the database, procurement, handling, uniformization of these structures. In many relevant studies in Structural Bioinformatics, several initiatives have been successful in developing tools for computational manipulation of biomolecules. Programming languages like Python, Perl and Java R are widely used to treat biomolecules, however they need great the effort to use as require a background in computer programming. In this sense, there is a growing demand for tools that can deal with these structures without the need for direct use of programming languages. Initiatives using strategies based on web services, facilitates access automated a set of specific tasks that often require a background in computer programming, on the other hand the use of tools that use graphical interface, to processing of the content when interact directly with users of the system eliminating the programming interface. In this study, we compared the main libraries developed with different languages for the specific purpose of treating biological data called here BIO :: , the statistical software R, and the tools, PDBest, Pro SA Web, PDB2PQR, PICES, Open Babel, and the repository RCSB, selecting a set of features that seemed more common to all of them to address the main basis of three-dimensional data on biomolecules, the PDB, or structures not deposited yet, but which maintain compatibility with the PDB format. Functions such as searching in this data base, ability to perform downloads in this data base, separation of the respective protein chains found in a file, atoms missing, residues missing, selecting a specific occupation for files with multiple occupancy, addition of hydrogen atoms at a given pH, removal of hydrogen atoms, re-enumeration atoms and residues containing structures numbering out of sequence, reproducibility of the experiments performed automatically, conversion of others biological file format and the need to background in computer programming for use of each tool. After this analysis, the PDBest tool showed better performance than the other solutions, offering a user friendly interface, quick responses and parallel processing capability.

Building LeifDB: a database for storing information about genome comparative analysis of the Leifsonia xyli

Pedro Bittencourt, Lucas Taniguti, Claudia Monteiro-Vitorello, Saul Leite, Wagner Arbex, Fernanda Almeida

Federal University of Juiz de Fora, ESALQ/USP

Abstract

In Brazil, sugarcane culture is used for the production of sugar and ethanol. However, there are diseases that affect this culture and cause considerable reduction in productivity. Amongst them is the ratoon stunting disease caused by the Leifsonia xyli subsp. xyli bacterium, which is one of the most onerous to the sugarcane sector. There are many species that compose the Leifsonia genus and only of the them is plant pathogenic, the others are free-living bacteria. This genus comprises seven species that have been isolated from distinct niches, such as plants, soil, distilled water, and an Antarctic pond. Leifsonia xyli is the only species causing plant disease and comprises of two subspecies: L. xyli subsp. cynodontis (Lxc), a pathogen that causes stunting in Bermuda grass (Cynodon dactylon) and L. xyli subsp. xyli (Lxx). The objective of this work is to organize the information about completely sequenced genome of these bacteria, when these genomes are available. Our goal is to establish an environment that organizes information about these bacteria in order to assist sequence comparison and functional annotation. In this sense, we report the construction of LeifDB, a relational database that groups and standardizes information about the genome sequences of Leifsonia species. Currently, the LeifDB contains the 11 completely sequenced genomes (i.e., five Clavibacter and six Leifsonia). Clavibacter species are plant pathogens and the closest of the Leifsonia species. The database contains the ORF (Open Reads Frame) prediction of these 11 genomes using Prokka software (`http://vicbioinformatics.com/`) and functional categorization based on the COG Database (`ftp://ftp.ncbi.nih.gov/pub/COG/COG2014/data`). We also stored the analysis generated by the OrthoMCL DB (Database of Ortholog Groups of Protein Sequence) with all the predicted proteins of the 11 species. The database is under construction and was created with the data management system MySQL. Next, we intend to propagate functional annotations of the Xanthomonas Genome Project to help the comparison with other genomes sequenced by the group in previous analyses. The main objective of this work is to organize a group of genes potentially associated with pathogenicity not present in free-living bacteria. With this work, we aim to provide researchers and other interested parties with a trusted, well-structured source to support their research and work.

Behavior of the major flaws over the last ten years in Protein Data Bank

Wellisson Gonçalves, Raquel C. de Melo-Minardi

UFMG

Abstract

Using PDBest (PDB Enhance Structure Toolkit), we show how some inconsistencies has behaved over ten years in the Protein Data Bank (PDB) database focusing mainly on the behavior of four problems encountered in the deposited structures in the database or who retain compatibility with the PDB format, missing atoms, missing residues, nonstandard residues and multiple occupancy. Despite being the most representative data base of biomolecules to study this structures, the Protein Data Bank (PDB) presents some problems that can skew and delay many computational analysis. There are two methods, more representative for obtaining this structures contained in the biomolecule files , X-ray crystallography (X-RAY), which represents 89 percent of the database and the Nuclear Magnetic Resonance (NMR), which represents 10 percent of the database. This amazing repository, receives new structures for almost 50 years and during this time, new equipment was created and older have been improved, new methods have emerged, showing better performance for obtaining the structures dealt with PDB files. Although the deposit process is increasingly cautious about the validation and review of new structures, there are two main factors contributing considerably to this inconsistencies in this database. The first of these are failures caused by technological limitations to obtain the structure and the data used in the validation process at the time of filing. The other point to consider, are the flaws that still persist after the validation process by unfavorable experimental conditions or persistent errors in the validation process. Using this tool, PDBest, it was possible to identify, with high throughput already at an early stage the files that could prevent or delay a possible experiment, also offering a detailed report with all the information processed, which can ease the deletion or treatment process structures in the new set of data.

Motifs Discovery Using Profile HMM and Evolutionary Algorithms

Jader M. Caldonazzo Garbelini, André Yoshiaki Kashiwabara Danilo Sipoli Sanches

Universidade Tecnológica Federal do Paraná

Abstract

The study on shared patterns of biological sequences called motifs, is important to help researchers understand how organisms work. Several inferences can be made through these patterns, as for example, function, secondary and tertiary structure of proteins and location promoters and regulatory regions of genes. Experimental methods, such as DNase Footprint, gel-shift, reporter construct assays, ChIP has been used to determine the position of motifs with relative success. However, find hundreds or thousands of likely candidate sites using experimental techniques demand a lot of time and money This makes computational methods of aligning an excellent way in search for motifs. They consist basically in perform multiple sequence alignment (MSA) of target sequences based on similarity of the residues. Has already been proven that the MSA is a computational complexity problem Non-Polynomial (NP). Problems of this class cannot be solved in polynomial time, i.e., depending on the number of input sequences, the alignment can be a prohibitive time. Although brute-force computational techniques can find always a optimum alignment, they cannot be employed because of the restrictions imposed by the MSA. Therefore, heuristic methods need to be applied. Profile Hidden Markov Model (PHMM) has been used with much success in Bioinformatics in the solution of the MSA. Although PHMM is a very robust method, in the course of its implementation, it can be pressure on local maximum or minimum. Evolutionary Algorithms (EAs) has been shown to be effective in solving large problems that have a complex search space. The EAs are non-deterministic algorithms, i.e., there is no certainty that they will find the optimal solution. However, you can find most of the time very close to the great solutions at relatively low computational time compared with mathematical programming techniques. In view of the above, this work is being developed a new methodology using PHMM and EAs, aiming to solve the problem of the MSA more efficiently.

Computational methods applied to identification of the Dairy Gir breed families

Gisele Silva, Tales Silva, Míria Bobó, Fernanda Almeida, Victor Menezes, Stênio Soares, João Cláudio Panetto, Wagner Arbex

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

Abstract

The Embrapa Dairy Cattle and the Brazilian Association of the Dairy Gir Breeders (Associação Brasileira dos Criadores de Gir Leiteiro/ABCGIL) develop for 30 years the Genetic Improvement National Program of the Dairy Gir (Programa Nacional de Melhoramento do Gir Leiteiro/PNMGL). Released in May 2015, the most recent PNMGL's sire summary evaluated and published the results about 300 males, using more than 30,000 information to their daughters or contemporary thereof and more over 32,000 records of dairy production of their daughters and contemporary. Among several sets of data, the PNMGL activities monitoring and maintains two databases (i) for store the certificate (registry) of the individuals (males and females) (ii) with the records of dairy production (females). This work will assists the PNMGL, because it will be important to find traits which can be decisive for increasing milk production. The database contains 92,488 animals registered with the following attributes: the IDs of the animal, father and mother; the animal gender and the animal birth date. The beginning of this study was a descriptive analysis of the database and create a graph where each node represents a animal and each edge represents a relationship between the animals and its parents. This graph was divided into connected components - which can be seen as "animal families" into the database - and a genetic inheritance was assigned according to affiliation of animals, within each of connected components. For the animals whose "genetic load" was supplied by a common ancestor, through father and mother, it is considered one with the biggest amount of transmitted genetic load. The descriptive analysis demonstrated that (a) 94% of animals are female and 6% are males; (b) about 22% of animals does not have a known parents; and (c) nearly 52% of animals do not have descendants. The graph was separated in 3,959 connected components; however, only 12 connected components have more than 10 animals. The largest component covers 94% of animals, containing 15 sires with more than 500 direct descendants. The largest descendants number of a single animal is 4,004, it means that about 4% of registered animals are direct descendant of the same sire. It was observed that 67% animals have some known genetic heritage in his grandparents ascending line. These results were the first steps in a project for explore and reveal new features about important parts of the dataset stored in the last 30 years by PNMGL.

API-Centric Data Integration for Human Genomics Reference Databases: Achievements, Lessons Learned and Challenges

J. S. Freitas, M. P. Caraciolo, V. M. Diniz, J. B. Oliveira

Genomika Diagnósticos

Abstract

Data Integration is a main challenge faced in clinical genetics where there are multiple heterogenous databases spanning several domains (e.g. biological, clinical variants and diseases) presented in confusing formats without clear and common standards. In variants analysis for molecular diagnostics applications, one central task is to connect biological information to clinical data such that specialists can determine the potential impact of that variant associated with a given disease. For this task, it requires the flexible assembly of tailored data sets continuously attention without wasting the biologists and geneticists time on searching several databases individually online, parsing, updating, cleaning and integrating those data in complex spreadsheets. To address this challenge, at Genomika Diagnósticos laboratory, we are building a platform that leverages Linked Data to provide integrated access to several bioinformatics databases such as OMIN, Clinvar, using a common and well-defined interface. With this platform, we expect to increase the productivity of our team of biologists, improving a assurance and quality control of our services. In this poster, we discuss how using linked data concepts can facilitate the execution of filters and query across several databases without knowing the particularities of each dataset. Exposing those datasets via Application Programming Interfaces (API's), it can facilitate the data access from several sources to a big data infrastructure, which provides integrated access to covering information about biological, carrier testing, variant analysis and literature mining. Additionally, we will present for the audience the challenges, experiences and achievements using our proposed architecture and how it can be replicated and extended for other sources.

EGene2 DB: automated pipeline construction system integrated with a database management system – application for the Photorhabdus luminescens MN7 genome project

Liliane Santana Oliveira, João Marcelo Pereira Alves, Carlos Eduardo Winter, Alan Mitchell Durham, Arthur Gruber

University of São Paulo

Abstract

EGene2 is an integrated system for building customized pipelines to automatically process and annotate biological sequences. In its current implementation, EGene2 comprises more than 50 components that can be combined to construct pipelines for customized and comprehensive genome annotation. These pipelines can include similarity searches, identification of protein motifs and domains, orthology classification, pathway mapping, and GO term assignment, among other tasks. In addition, EGene2 can generate report files in common formats such as feature tables, GFF3, and HTML. Originally, EGene2 only used direct manipulation of XML files for storing sequence information and annotation results. However, while XML files are very easy to use, their hierarchical structures are not adequate for data mining. Any new query requires a new piece of code to analyze the annotation files. Also, a large number of queries would be very time consuming. In this work we report the extension of the EGene2 platform to include a PostgreSQL persistence module. The original XML data model implemented an abstract representation of analysis results. During the development of the new database persistence module, the similarity between our original data model and Chado, a generic schema representing many of the general classes of data frequently used in modern biology, became clear. We have thus extended Chado to include information on pre-processing and to enable a curation system, and used it to implement the new PostgreSQL persistence module. This new module enables the use of the SQL language to manipulate much larger datasets and to perform a wide range of queries of the data. Simple queries, like retrieving sequences from a specific organism, to more complex queries, like the places in a genome where repetitive regions within specific size boundaries are located, are now possible. EGene2 DB was used to automatically annotate the draft bacterial genome of Photorhabdus luminescens MN7, a symbiont of entomopathogenic nematodes of the genus Heterorhabditis. The unusual tripartite interaction involving the bacterium, a nematode, and an insect constitutes an interesting model for the study of host-pathogen interactions. The integration of an automated annotation system and a relational database allows biologists to perform complex queries. This is a key aspect required for the identification of bacterial genes involved in the production of secondary metabolites and toxins that mediate symbiotic relationships with the nematode partner, the insect host, and free-living nematode models. Annotation results obtained with different queries will be presented. Support: FAPESP and CNPq

PROCESS DESIGN AND CONFORMITY ANALYSIS OF COMPUTATIONAL STRATEGIES FOR MOLECULAR DOCKING

Miller Biazus, Eduardo Spieler, Lucineia Heloisa Thom, Marcio Dorn

Federal University of Rio Grande do Sul

Abstract

The designing of particular processes with specialized vocabulary and dynamic characteristics such as those encompassed by existent computational strategies for molecular docking are very complex, not only because of the variety of existent computational strategies, but also because they require the knowledge of very specific domain terms which can lead to interpretation problems, ambiguities and misunderstandings between the process stakeholders. The molecular docking process includes macro-activities such as receptor and ligand preparation, solvent adding and binding site identification which can be simulated by several and different tools (e.g. DockThor, Autodock, GOLD, Glide). However, each tool has its own techniques for performing the macro activities which implies different molecular docking processes in particular when compared with the process proposed by the molecular biology literature. In this approach we propose a literature-based design of the molecular docking process using the standardized graphical notation BPMN (Business Process Modeling Notation). From the molecular docking process model we extract a corresponding propositional logical-based meta-algorithm. This algorithm can serve as a reference for the development and improvement of tools related to molecular docking. Subsequently, we analyze a set of tools for molecular docking focusing on their computational strategy process. Our goal is to identify similarities and deviations of their process when compared with the molecular docking process obtained from the literature. As main contribution our approach provides a more standardized and systematic view of the molecular docking process which can be very useful for learning and to assist the development of more accurate tools for molecular docking. The comparative analyses of the molecular docking techniques can potentially help users to choose the technique which best matches their needs. Acknowledgement:This work was partially supported by grants from FAPERGS (002021-25.51/13) and MCT/CNPq (473692/2013-9).

Koala: a web-based platform for protein structure evaluation and analysis

Alexandre Defelicibus, Rodrigo Faccioli, Alexandre Delbem

Universidade de São Paulo

Abstract

Biomedical researches have been generated a large amount of data. As the generation and manipulation of high volume of data has become more accessible, the researchers have faced with challenge in order to analyze these data. There are many computation solutions, but Galaxy project has been highlighting by the increasing number of users and researchers. More specially, we have developed a server to analyze and share the data from multidisciplinary problem: protein structure prediction (PSP). The PSP problem consists in finding the tertiary structure of a protein from its amino acid sequence. The folding process is not fully understood, it has been modeled as an optimization problem. Evolutionary Algorithms (EAs) are based on the Darwin's Evolution Theory and one of the optimization techniques applied to PSP, a problem whereas have many local optima and large search space. These features are able to heuristic-based approaches. In order to challenge in analyze large amount of data, we have developed Koala, an web-based platform. The proposed server is composed by several tools for protein structure evaluation and analysis. Currently, Koala has more than ten ab initio algorithms with different approaches and heuristics to predict tertiary protein structure. In terms of analysis, the mostly used algorithms to assess protein structure similarity Root Mean Square Deviation (RMSD), Global Distance Test (GDT) and Template Modeling Score (TM-score) are available on the server. Besides, this server provides our own methodology of analysis that are not parameter dependence, such as protein size, type and reference structure. Koala is a web-based platform to protein structure prediction with an easy-to-use user interface under cloud environment. One important feature is users can build their own workflow from the available tools, which automates the data analysis. Furthermore, all results from Koala can be shared and reproduced by other users, which guarantees the reproducibility of research.

A system for gene annotation of the Copaíba (Copaifera multijuga)

Andressa Rodrigues Alves Galvao Valadares, Waldeyr Mendes Cordeiro Silva, Maria Emilia Machado Telles Walter, Maristela Terto Holanda, Marcelo Macedo Brigido

Universidade de Brasília

Abstract

The resin-oil produced by the Copaíba trees, used as a defense mechanism against its predators, has been used in traditional medicine over 500 years. It's pharmaceutical properties have been scientifically proven, such as anti-microbial, anti-inflammatory and antineoplastic activities. Through the Copaiba's transcriptome, the metabolic pathways could be determined, particularly those involved in the terpenoid production, the main oil components. In this work, we have modeled and implemented a relational database to support the set of annotations from the main databases available, e.g. Uniprot (SwissProt, Trembl) and BRENDA, besides manual annotations. A relational database is a collection of data items organized as a set of formally-described tables from which data can be accessed or reassembled in many different ways without the need to reorganize the database tables. A total of 62.839 contigs are available in our database, where 5.373 contigs were annotated as enzymes related to secondary metabolism. In total, 383 different enzymes of the secondary metabolism could be identified. The database stores the Copaíba data and provides access to all generated data concerning the classification of secondary metabolic pathways. Additionally, it supports new annotations using a web interface, enabling collaborators to annotate contigs manually. Furthermore, it is possible to search and compare all types of contig annotations in the database. This database and interface implementation supports scientific collaborations among the geographically dispersed institutions involved in the Copaíba project, making it a tool that will help elucidate new pathways of secondary metabolism of Copaifera multijuga improving the knowledge on this plant.

CREATING AND STRUCTURING A DATABASE OF CRY GENE FAMILIES FROM BACILLUS THURINGIENSIS AND IMPLEMENTING AN IDENTIFICATION TOOL

Erinaldo Nascimento, Laurival Vilas-Boas, Kátia Gonçalves, Gislayne Vilas-Bôas, Alessandro Bovo

Federal University of Technology - Paraná, Cornélio Procópio, Brazil, Universidade Estadual de Londrina, Federal University of Technology - Paraná, Londrina, Brazil

Abstract

Bacillus thuringiensis (Bt) is a spore-forming bacterium that produces the Cry toxin as parasporal crystals, which has been shown to be effective in the control of agricultural pests and vector mosquitoes, as well as in the development of transgenic plants resistant to insects. The adoption of Bt biopesticide has resulted in a significant reduction in the use of chemical insecticides, since it makes a biological control of insects, with no risk or harm to human health. However, many insect pests are not susceptible to the Cry toxins identified so far. Up to now, more than 700 cry genes from at least 70 groups have been identified. An alternative to pests that are not susceptible to parental Cry toxins is the isolation of other Bt strains with new cry genes with enhanced toxicity, as well as the identification of the receptor molecules and binding epitopes, screening of new Cry proteins with new insects and specific receptors. Another alternative is the in vitro genetic evolution of such toxins. The objectives of this study are the creation and structuring of a curated database of cry genes from Bt and the design and implementation of a tool for the classification and identification of the cry gene family of a given isolated nucleotide sequence or set of contigs. Therefore, this study will support the activity of carrying out a phylogenetic analysis and classification of cry gene groups, as well as the optimization to predict susceptible targets to Cry toxins. The process of analysis and phylogenetic classification will use data mining and pattern recognition methods with biological data. The computational tool will consist of a strategy for the discovery of likely candidates to be a cry gene based on score, and a characterization method that makes a phylogenetic analysis based on similarity. The output of the tool will be the sequence alignment and phylogenetic tree of the likely cry gene. The key issue is to provide a web interface and an online database to support the biologist in the task of classification and identification of Cry proteins, based on the DNA sequence, aiming at the description of biotechnological products with enhanced toxicity that may act on the control of insect pests and disease vectors in a broader spectrum of action reaching those currently not susceptible to Cry toxins or that are inefficiently controlled.

DataPGx: making pharmacogenetic research more productive

Welber Oliveira, Luiz Alexandre Magno, Rosangela Hickson

Faculdade Infórium de Tecnologia e INCT de Medicina Molecular/Fac. de Medicina - UFMG,

Abstract

Research data should be maintained, and their use optimized. However, in many laboratories, there is no guideline of regarding the need to preserve and document pharmacogenetic (PGx) data. Since middle 1990s, most scientists eventually have been working with data stored in spreadsheet software, such as Microsoft® Excel®. Although much of the computer technology has changed in the past years, most PGx laboratories remain keeping data in incomplete, unclear or multiple spreadsheets. This approach has been demonstrated to be prone to errors, difficult to use in groups, and slow the data input/output significantly. In fact, computer scientists advise spreadsheets should never be used as a database. Self-experience has showed us how difficult is to find data that were created years or decades ago, but is needed now. Here we provide a user-friendly, customized data storage platform that will make easier to store PGx data. We call it DataPGx.: Laboratories would greatly benefit from this online database solution. We believe it will cover the current demands on recording, keeping, analyzing, sharing, and accessing the data anywhere - to closer cooperation between scientists and ensure long-term storage of scientific data. DataPGx is centrally managed with unparalleled security features, flexible workflow engine and robust audit trails capability. For example, data can be discovered and accessed through an object's detailed attributes such as creation date, author, keywords, project, study, grant, and more. All this data can reside on very different, incompatible platforms crossing multiple administrative domains, but tied together under a single laboratory or research group. This system can also track and ensure data provenance and data reproducibility, and control data access – exactly what is needed to manage and protect scientific data from patients. With these tools available to scientists, we have the ultimate aim of making PGx research more productive.

Development of new models of proteins network clustering for transcriptogram methodology

Alex Augusto Biazotti, André Luiz Molan, Agnes Alessandra Sekijima Takeda, José Luiz Rybarczyk Filho

Universidade Estadual Paulista - UNESP

Abstract

As technologies evolves, the studies in gene expression have been increasing. These analysis results in a large amount of data and new techniques are essential to understand the cellular processes involved in biological responses. In this work, we propose the development of two new protein network clustering techniques based on transcriptogram model [1,2]. This methodology use a cost function and the montecarlo method for minimize the cost function. The network is converted in a matrix, and each matrix element is rated in function of their neighborhood. In the transcriptogram methodology, the first neighbors in the left, right, up and down for each matrix element is evaluated, defining this model type as ”cross”. The new ”ring” model evaluates all first neighbors around each matrix element, and the ”X” model evaluate the first neighbors in up-left, up-right, down-left and down-right for each matrix element. We applied the models to networks of Saccharomyces cerevisiae, Aedes aegypti which were obtained from STRING database (confidence score > 0.7) and Homo sapiens from STRING and STITCH (confidence score > 0.7) databases. The H. sapiens network has 91647 nodes (proteins ans small molecules) and more than 800 thousands of interactions, while A. aegypti (7563 proteins and 147606 interactions) and S. cerevisiae (4655 proteins and 47415 interactions). The first evaluation was the processing time of each model in 5000 montecarlo steps, in this case, the algorithm for ”cross” model is faster than transcriptogrammer [2], and allows users to submit large networks (> 50000 nodes). The second evaluation was the clusters identification. The ”ring” model was superior when compared to the others to evidence the proteins clusters. We found 7 major clusters for H. sapiens, 8 cluster for S. cerevisiae, 6 clusters A. aegypti. The clusters analysis for these three networks indicated the ”ring” model as superior when compared to the others. The next step will be the application of the clustering method in H. sapiens networks with miRNA, lncRNA, proteins, small molecules in the cancer problem.

CoGA: an R package for differential co-expression analysis based on network spectral and structural properties

Suzana de Siqueira Santos, Thais Fernanda de Almeida Galatro, Rodrigo Akira Watanabe, Sueli Mieko Oba-Shinjo, Suely Kazue Nagahashi Marie, André Fujita

Institute of Mathematics and Statistics, University of São Paulo, School of Medicine, University of São Paulo

Abstract

Gene expression analysis plays an important role in the identification of genes associated with a disease. One of the most common approaches for differential expression analysis is the test of the equality in the average expression of single genes between two phenotypes. An alternative to that standard gene expression analysis is the gene set analysis, which tests the differential expression of sets of functionally related genes. That approach has been very successful by enhancing statistical power and aggregating prior biological information about the gene sets. One limitation of gene set expression analysis is that it does not take into account the correlation structure among the genes expression levels, which is also known as gene co-expression graph (network). To address that limitation, we developed statistical methods to test the equality of the gene co-expression graph structure between two phenotypes. We implemented those methods as an R package with a graphical interface called CoGA (Co-expression Graph Analyzer), available at `http://www.ime.usp.br/~suzana/coga/`. The CoGA tests are based on Information Theory concepts, such as the Shannon Entropy and the Jensen-Shannon divergence, applied to the distribution of the graph spectrum (set of eigenvalues of the graph adjacency matrix). The package also includes statistical tests for graph measures that are commonly used to analyze real networks, namely degree, betweenness, closeness, and eigenvector centralities, degree distribution, shortest path length, and clustering coefficient. Besides the tests for the graph structural properties, other available features are gene expression and co-expression plots, ranking of genes and gene pairs according to their "importance" in the network and the standard differential gene expression analysis. Our simulation experiments and analyses of real datasets suggest that the tests available at CoGA effectively control the rate of false positive and might be useful in the identification of genes that play a key role in a disease. Funding: This work was supported by FAPESP (grants 2011/50761-2, 2012/25417-9, 2013/03447-6, and 2014/09576-5), CNPq (grants 304020/2013-3 and 473063/2013-1), CAPES, and NAP-eScience-PRP-USP.

A highlymutation viruses genomic analysis system: highlighting the HIV sequence distribution

José Irahe Kasprzykowski, Felipe Guimarães Torres, Beatriz Abreu Gomes, Artur Trancoso Lopo de Queiroz

Universisdade Estadual de Feira de Santana, Fiocruz-BA

Abstract

Viruses infections are major public health problems. High mutation rate viruses increase management difficulty. HIV only infects 40 million people worldwide and is considered by the World Health Organization a large scale pandemic, with no actual cure. New information regarding etiological agent is necessary. The present data analysis could help on new therapy and vaccine development. However, the dataset is vast, over 500,000 sequences available on GenBank, and this data lacks subtyping and genome location. These information are essential for epidemic control procedures and new treatment development. To help minimize these problems we developed a system for automated indexing and subtyping from GenBank data. The tool performs sequence map according to HXB2 and subtyping by comparison with subtype reference sequences. The alignment performs a modified NeedlemanWusch algorithm, with the Gotoh affine gap penalty implementation. All 582,678 sequences were mapped in 5 days and 14 hours and subtyped in 1 day and 7 hours with our algorithm, while the original approach was estimated to finish in 97 years. Our tool was able to analyse the massive data in a reliable time. No current subtyping tool can analyse this highthroughput data. Our results showed that p ol and g ag genes were the most prevalent genes on the dataset, and could be explained because treatment and subtyping are based on these genes. Moreover, the structural genes were most prevalent, with 66.41%. This highlighted the low representation of regulatory genes on available data. The subtyping results showed that the subtype B was most frequent, with 45.96%. The recombinants together represent 43.37%. This shows an increase on recombinant prevalence, compared to Los Alamos dataset (∼320,000 sequences). Furthermore, subtype C presented only 4.12% and the other pure subtypes less than 4%. Also, the geographical data was taken into account and USA presented higher frequency, with 24.50%, showing a significant country bias. Our results present a new HIV subtype distribution with the most complete and recent dataset. Herein, we presented a new user friendly software for massive data analysis of viruses. This software is able to analyse highly mutational virus data, such as HCV and HIV in reliable time. Further, severe country bias raises questions regarding world subtype distribution. The analysis of all sequences from HIV provides new epidemy insights about subtypes and country distribution.

ClustEval - A Fully Automated Cluster Analysis Framework

Richard Röttger, Christian Wiwie, Jan Baumbach

University of Southern Denmark

Abstract

Equipped with sophisticated biochemical measurement techniques, a tremendous amount of biological data of ever increasing precision and quality is produced every day. As welcome as this trend is, it also poses a multitude of problems. In order to cope with this amount of data, automatic knowledge extraction techniques are utilized. One of the most popular methods of automated knowledge extraction is the so-called clustering, i.e., the grouping of similar objects into clusters. Even though clustering is a long standing problem in computer science, conducting a high-quality cluster analysis is all but straight forward. For the practitioner the very plethora of existing clustering algorithms is already a huge obstacle. Every newly proposed method is compared only to a selected handful of already existing tools on a possibly biased selection of datasets. Almost every clustering tool uses its very own input and output format which renders the conduction of large comparative clustering studies a very exhausting and error-prone process. Furthermore, every clustering tool requires the user to set at least one parameter, basically defining whether the result should be few large clusters or many small clusters. Finding the optimal parameter is again a very tiresome and error-prone process which often requires in-depth knowledge of the algorithm itself. With ClustEval, we introduce an integrated clustering framework, assisting the user in all steps of cluster analyses, from data preprocessing and parameter optimization to evaluating the reported clusters. The flexibility of the framework allows convenient extension with new tools, datasets, and quality measures. Furthermore, the layman is able to inform himself about the different clustering tools and their performance upon different types of datasets on an easy to use website. All this information is based on almost 4 million cluster validity indices which also allowed us to deduct a guideline for applying clustering tools in a biomedical context.

Genomic annotation of Leishmania braziliensis and storage data on data model

Felipe Torres, José Gonçalves, Beatriz Abreu, Vinícius Coutinho, Artur Queiróz

Centro de Pesquisa Gonçalo Moniz - FIOCRUZ

Abstract

The cutaneous leishmania is infectious disease that affect about 12 millions of people around the world. The main ethological agent this disease on Brazil is the Leishmania braziliensis. But the genome annotation this specie is bad and very inaccurate. The accuracy is lower because exist many hypothetical genes and putative genes on annotation published in NCBI and TrytripDB. This work made the genome annotation process again. Was used new techniques of genomic annotation on annotation process. Was obtained predicted genes and predicted non coding RNA using GLIMMER and GENSCAN. The predicted genes was identified by similarity using similarity search algorithms. For identification, we used the SWISSPROT database and the BLASTx software for compare proteins and predicted genes. After the identification, was identified the protein function using the software AmiGO. Was identified 2382 genes resumed in 1245 proteins. All founded genes was mapped and aligned on Leishmania braziliensis genome Was developed a data model adapted for generated data in MySQL. The generated data was storaged on a database model developed. Before the next step, we tested the database model for to search performance problems on model. The new database is LeishDB, implemented on MySQL and your frontend system was developer on PHP with Javascript. For visualize the genomic data, we are using the genome browser, JBrowser on frontend. The genome browser need to be configured and adapted for project. For it, we made BED, SAM and GBK files with data of genomic annotation. Those files were created to using Python Scripts and BioPython framework. Using the three files was created the JSON database for JBrowser. This final database can be accessed by the address: `http://www.leishdb.com`

High throughput sequence subtyping tool for highly-mutation viruses

José Irahe Kasprzykowski, Felipe Guimarães Torres, Beatriz Abreu Gomes, Artur Trancoso Lopo de Queiroz

Universisdade Estadual de Feira de Santana, Fiocruz-BA

Abstract

Virus epidemics represent a major health problem. To study and surveillance purposes the nucleotide sequences of this individuals are sequenced and stored on global databases. Nevertheless highlymutation viruses present a challenge on surveillance for it has an large amount of data and various subtypes. About this kind of organism, the main global biological databases still lack of subtyping data. It also occurs due to the amount of data and computational complexity involved in the subtyping process. To execute this process the sequence must be aligned to reference sequence of all subtypes identified. This process is commonly executed by an heuristic algorithm tool such as BLAST. Notwithstanding, this methodology lacks of precision, once not always the alignment returned is the optimal one. Likewise, the heuristic methodology is directly dependent of the data arrangement and condition. Thus the solution is to apply exhaustive mathematical algorithms which are not data dependent and always return the best alignment possible. However the optimal algorithms such classical Smith & Waterman are time and resource consuming. As an example, to subtype all sequences available from HIV-1 on GenBank, the classic approach would take around 97 years to finish. To help solve this problem, we developed a tool for subtyping high amount of sequence in a short period of time. For this we used a modified version of Smith & Waterman algorithm. We optimized this algorithm by reducing machine resource use, replacing the main matrix for two dynamic vectors. We also parallelized some algorithm activities. Along with algorithm modification we also created an strategy to reduce the alignment amount needed for the subtyping process. This strategy consist in grouping the subtype reference sequences by recombination derivation which allows the reference groups creation. This reduces the alignment amount needed by comparing the sequences first to the reference group and then to its members. We also reduced this amount by applying an summary withdrawal strategy. This strategy consists in arrange the reference sequences by potential score, and stop the subtyping if the last score is greater than the next potential score. With this algorithm approach we were able to reduce the time needed to subtype all HIV-1 sequences to only 2 day and 5 hours. Applying the optimized strategies we reduced this time by $\sim$50

Analogous Enzyme Resource

Alexander Franca, Marcos Catanho, Ana Carolina Guimarães

Instituto Oswaldo Cruz

Abstract

Abstract Since enzymes control almost all biochemical reactions in the metabolism of living organisms, it is extremely important to characterize the genes encoding enzymatic activities. The most successful approaches to perform this task are based on homologous searching, using local alignment algorithms to find statistically significant similar sequences in protein sequence databases. Comparisons of metabolic pathways computationally predicted in completely sequenced genomes of diverse organisms evealed incomplete pathways or even absent enzymes in some steps of important metabolic pathways. In several cases the ”missing” enzymes were ”substituted” by functional quivalent molecules, able to catalyze the same reaction but exhibiting no significant similarity between them, herefore escaping detection by methods based on homologous earching. These alternative forms, known as analogous enzymes, arise from independent evolutionary events, converging for the ame biological function. Several studies suggest that the fraction of enzymatic activities in which multiple events of independent origin have occurred during evolution is substantial. However, this subject is still poorly understood, and a comprehensive investigation of the occurrence, distribution and implications of these events, involving organisms whose genomes have been completely sequenced, has not been accomplished so far. Fundamental questions such as ”how analogous enzymes originate?”, ”why so many events of independent origin have apparently occurred during evolution?”, and ”what are the reasons for the coexistence in the same organism of distinct enzymatic forms?” remain unanswered. In this work, we describe the development of a computational system/resource designed to assist analyses regarding sequence, structure and evolution of analogous enzymes in completely sequenced genomes of organisms representing the three domains of life, as well as to assist the computational prediction and metabolic reconstruction of analogous enzymes in protein datasets provided by users. The core of this system consists of an updated version of the computational pipeline developed by our group to predict putative analogous enzymes (AnEnPi) employing protein sequences available in public databases. Browsing a user-friendly web interface, users will be able to analyze data such as (i) global alignments and HMM-profiles representing groups of similar enzymatic forms, (ii) domain, folding and 3D structure information that characterize alternative enzymatic forms, (iii) distribution of analogous enzymes according to enzymatic activity, metabolic pathway, and taxonomic unit (phylogenetic pattern), both globally or for selected groups of analogous forms, enzymatic activities, metabolic pathways and/or organisms. Financial Support CAPES, PAPES-FIOCRUZ, CNPq, FAPERJ, and Plataforma de Bioinformática Fiocruz RPT04-A/RJ

COMBINING TEXT AND CONTENT BASED IMAGE RETRIEVAL ON LARGE MEDICAL RESOURCE DATABASES

David Silva Guedes, Thiago Antônio Teixeira Lima, Katia Cristina Lage dos Santos, Tiago Silva de Bessa, Diego Moreno Trepim

Instituto Metodista Granbery

Abstract

The advance of information technologies applied to medical area has turned possible to obtain non-explicit knowledge from large databases. As a result, it becomes necessary to develop tools and techniques more effective and efficient to improve the analysis of large volumes of data. In this context it is necessary to develop a Picture Archiving and Communication System (PACS) that aim to combine text annotations and radiological images to define medical diagnostics. The related literature indicates that a quality model using Content Based Image Retrieval (CBIR) and textual evidences applied to PACS is one of the major gaps to be explored in the area. CBIR works on the image features as the color, texture, shape and edge or the combination of them. To implement a CBIR system is necessary an algorithm to extract this feature and a similarity distance method. This two algorithms defines an image descriptor. Based on this context this study aims to evaluate different strategies for medical image retrieval, combining CBIR and textual evidences added manually by professionals involved in the diagnosis and prognosis of diseases. To implements the PACS it was chosen the cross platform language Java and two free libraries: Lire and Apache Lucene. At first it was carried out the evaluation of three image descriptors implemented in the library Lire: CEDD (color and edge), Tamura (texture) and Edge Histogram (edge). The choice of these image descriptors stems from the nature of the medical images obtained by radiography, usually monochromatic and with different resolutions. The aspects evaluated were the quality of the first images returned (top images) and the execution time necessary to generate the content information data sets using the training images. The former experiments were performed with 1,000 real radiography medical images from ImageClef 2009 database, with a total size of 71.6 MB. Preliminary results showed that considering the first six images (top 6) the descriptor CEDD obtained the best result. This result considered the average distance between the query image and the top 6 images returned. In terms of execution time to extract the content characteristics the Tamura descriptor takes twice as much time than CEDD and Edge Histogram descriptors. The experiments were performed on a machine with the following settings: 6GB RAM memory, Intel Core i5-2520M 2.50GHz, Windows 8.1 Pro x64. The next step is to develop a mechanism that performs the combination the content information and textual evidences using Apache Lucene.

LZ78 factorization using the FM-Index

Daniel Nunes, Felipe Louza, Guilherme Telles, Mauricio Ayala-Rincón

Universidade de Brasília, Universidade Estatual de Campinas

Abstract

Lempel-Ziv factorization plays a fundamental role in data compression and text indexing. There are different Lempel-Ziv factorization approaches. We are interested in the algorithm of 1978 known as LZ78. The LZ78 compression algorithm scans the text from left to right replacing substrings in the text by a pointer to a previous maximal occurrence of a "factor". Formally, let $S = S[1, n]$ be a text of length n over an alphabet Σ. The LZ78 factorization of S is a sequence $f_1 \cdot f_2 \cdots f_m$ of factors of S, such that each f_i is either a letter $c \in \Sigma$ that does not occur in $f_1 \cdot f_2 \cdots f_{i-1}$ or is $f_j \cdot c$, where f_j is a previous factor ($1 \leq j < i$) that is the longest prefix of $S[|f_1 \cdot f_2 \cdots f_{i-1}| + 1, n]$. The LZ78 factorization can be encoded by a sequence of pairs $b_i = \langle j, c \rangle$, where $f_i = f_j \cdot c$ ($j = 0$ when $f_i = c$). For instance, let $S = aabbbaababab\$$, the LZ78 factorization of S may be encoded as $\langle 0, a\rangle\langle 1, b\rangle\langle 0, b\rangle\langle 3, a\rangle\langle 2, a\rangle\langle 4, b\rangle\langle 0, \$\rangle$. The LZ78 factorization can be easily computed using *tries*, a well-known data structure used to index strings. In this work we show how to compute the LZ78 factorization using the FM-index together with the segment-tree. The proposed algorithm constructs the FM-index for the reversed string S^{rev}, performing a backward search for each symbol $S[i]$ helped by the segment-tree and finding the longest previous factor starting at position i of S. The algorithm is being implemented and validated through comparative tests against results of related works. The main goal is to reduce the memory usage of the existing solution. This algorithm may be applied in the compression of huge databases, such as genomic and natural language repositories.

Two different ways of obtaining sequences to train hidden Markov models for searching transposable elements

Victor Campos, Victor Barella, Carlos Fischer

São Paulo State University (UNESP), Universidade de São Paulo University (USP)

Abstract

Transposable Elements (TEs) are nucleotide (NT) sequences capable of ”moving” within a genome. There is great biological interest in TE identification and classification. The focus here is on the class LTR-Retrotransposons (superfamilies Bel-Pao, Copia and Gypsy). Profile Hidden Markov Models (HMMs) are models used to recognize pattern within sequences and profile, e.g. TE superfamilies. HMMER is a widely used application for searches in NT sequences. HMMER uses sequence alignments for training HMMs. Repbase, a database with a large amount of TE information, can provide two types of NT sequences: internal portions of TEs and coding regions (CDS) of TEs. However, only a part of these CDS have portions related to conserved protein domains characteristics of the superfamilies of interest. Considering Repbase, the traditional way (protocol) of generating alignments to train HMMs consists of using complete internal portions of TE. Here, we present two alternatives to this traditional protocol. The three tested protocols provide sets of representative sequences compound of: (i) complete internal sequences; (ii) all CDS sequences; (iii) CDS sequences that have at least one conserved domain. These protocols are named: Complete, CDS and CDS-domain, respectively. These sets of sequences were used to train three independent HMMs. Using HMMER, these HMMs were run on data from Repbase (performing a k-fold cross-validation method) and on the Drosophila melanogaster genome. We used precision, recall and F-measure to evaluate and compare the three protocols. Data from Repbase: filtering the e-values returned by HMMER, within range 1e-05 to 1e-34, the results show that the best protocols (with F-measure/filter) for each superfamily are: (i) Bel-Pao: Complete (0.985/1e-12); CDS-domain (0.983/1e-11); CDS (0.958/1e-10); (ii) Copia: CDS-domain (0.981/1e-13); CDS (0.981/1e-11); Complete (0.980/1e-18); (iii) Gypsy: CDS-domain (0.950/1e-15); Complete (0.937/1e-15); CDS (0.588/1e-08). Using D. melanogaster genome: considering the three protocols and filters defined using Repbase, the best protocols (with F-measure) for each superfamily are: (i) Bel-Pao: CDS (0.851); CDS-domain (0.729); Complete (0.728); (ii) Copia: CDS-domain (0.955); Complete (0.952); CDS (0.755); (iii) Gypsy: CDS-domain (0.945); Complete (0.889); CDS (0.723). The results showed that, considering data from Repbase, the three protocols presented the same performance, except CDS in Gypsy. On a real genome, compared to protocol Complete (the traditional one), the CDS was better for Bel-Pao and, for Gypsy, the CDS-domain presented better performance. Results also show that the HMMs could be used together to increase the number of correct predictions, particularly for Gypsy.

Computer simulation model development to evaluate quantification methods for gene expression by RT-qPCR

Carlos Diego de Andrade Ferreira, Leandra Linhares Lacerda, Priscilla de Barros Rossetto, Sandro Leonardo Martins Sperandei, Marcelo Ribeiro-Alves

FIOCRUZ - Instituto Nacional de Infectologia Evandro Chagas -, FIOCRUZ - Instituto Oswaldo Cruz, UFRJ - Instituto de Biologia (Genética), FIOCRUZ - Instituto de Comunicação e Informação Científica e Tecnológica em Saúde

Abstract

Real-time RT-PCR (RT-qPCR) is currently the gold-standard technique for estimating relative gene expression. Nonetheless, quantification methods that assume constant amplification efficiency and/or are unstable to variations encountered in real-time amplification kinetics can introduce considerable errors on the results of RT-qPCR. A computer simulation model formulation for RT-qPCR reactions that consider various combinations of systemic factors known to influence the amplification kinetics could be useful either to evaluate the currently quantification methods or to development of new and more precise ones. Therefore, we aimed to formulate such a model from empirical observations where known systemic variations were introduced: different RNA integrity number (RIN) samples, amount of dNTPs and primers, and amplification efficiency. We used HeLa cells cultures to generate empirical qRT-PCR amplifications on different experimental conditions. The total RNA extraction of HeLa cells culture was made by RNeasy kit (Qiagen) followed by Bioanalyzer assays (Agilent Technologies) for quantification and quality evaluation. Synthesis of cDNA was performed by SuperScript III (Thermo Fisher Scientific) and the qRT-PCR reactions were conducted in a 7500 Fast Real-Time PCR detection system (Applied Biossystem) with SYBR® Green, using the following experimental conditions: samples with RIN 7 or 10; optimal amount of dNTPs or 75% of optimal; optimal amount of primers or 90% of optimal; low (HPRT1; eff=1.83) and high amplification efficiency (B2M; eff=2.03); and concentrated and 1000 times diluted cDNA . Overall, we conducted 192 reactions in 32 distinct experimental conditions in hexaplicates. Amplification efficiencies for HPRT1 and B2M genes were estimated by serial dilution method and optimal amount of dNTPs and primers were set on previous assays. The formulation of the stochastic simulation model introduced as 8 main parameters: RIN; amount of primers; amount of dNTPs; amount of primer-dimmer formation; amount of initial transcript cDNA; amplification efficiency; noising factor; and, bleaching factor. The global Mean Square Error (MSE) between simulations and empirical observations among different experimental conditions was 0.012 (CI95%= -0.011-0.035). Moreover, when we evaluated all variation regarding the different conditions between the simulated/used cDNA concentrated we observed a MSE of 0.0232 (CI95%= -0.0223-0.0689) and 0.00035 (CI95%= -0.00034-0.00102) when simulating/using 1000x diluted cDNA. Our preliminary results show a good adhesion between the simulated RT-qPCR kinetics and the empirical observations. We will further generate benchmark datasets from the simulation model to evaluate the currently quantification methods and defined guidelines with the most appropriate methods of quantification for the diverse amplification conditions.

Strategies and best practices for automated benchmarking on multiple cancer/germline variant callers

Marcel Caraciolo, Victor Monteiro

Genomika Diagnósticos

Abstract

As novel nextgen-sequencing applications emerge, the number of bioinformatics tools continues to grow. In the context of variant calling pipelines, the options for bioinformatics softwares for each step of the workflow is huge. The researchers want to evaluate these softwares in the pipeline and check if it will work best in the analysis observing several parameters such as: performance, accuracy and formats for inputs and outputs. Therefore, in order to test these new tools, the researchers must install it manually and compare the pipeline results against others' using test data. Naturally, they willl perform several benchmarks, which can lead to a subjective comparison, sometimes even lacking reproducibility. Combining this with the current large numbers of bioinformatic tools published, it is difficult and demmanding for researchers to objectively evaluate which software will work best in the analysis using identical data and identical settings for the running tools. To mitigate these issues, we developed an automated benchmark suite, which places the desired tools in the pipeline and performs all the benchmarks against test data. This allow softwares to be evaluated simultaneously and without requiring manual installation or setting of parameters. The bioinformatics pipeline with the selected softwares and specific versions can thus constantly be evaluated against current existing datasets and the results of the benchmarking then can be shared with other researchers as they are generated. The goal of this poster is to show with examples, using our developed benchmark suite, how we evaluated cancer/germiline variant calling to compare multiple variant callers analyzing their specificity/sensitivity and improvements in speed. We believe that our experiences can be quite helpful for the bioinformatics community in order to improve their strategies on evaluating variant analysis pipelines.

A GPU-based algorithm to calculate k-mer frequency

Fabrício Vilasbôas, Carla Osthoff, Oswaldo Trelles, Kary Ocaña, Ana Tereza Vasconcelos

LNCC - Laboratório Nacional de Computação Científica, UMA - Universidad de Málaga

Abstract

Improvements in technologies as next-generation sequencing (NGS) or high-performance computing (HPC) contribute to expanding researches in novel domains of bioinformatics. Metagenomics is an application area which attend to explore bacterial environment. Classic metagenomics experiments related to the calculation of k-mer frequency can perform days or weeks if executed in standalone execution. The reduction of the time and processors consuming is needed for metagenomics experiments. The calculation of k-mer frequency algorithm that provides the best performance is the Jellyfish, a multi-threaded algorithm based on multicore architectures. Recent mprovements on GPU's technologies provides the possibility to implement algorithms to calculate the k-mer frequency on manycore architectures at a low price. The GPU architecture contains up to thousands of cores and can associate one thread per nucleotide on each read providing the ossibility to perform all k-mers calculations at the same time, providing up to one order of magnitude of performance gain. This work presents a GPU-based algorithm that calculates k-mer frequencies to be used for decreasing the data mining processing time on GPU's based metagenomics applications, such as k-means classification. To validate our approach, we use a set of human metagenomes obtained from NCBI database and compared to the classic OpenMP-based algorithm, Jellyfish. We evaluate the performance from Jellyfish and GPU algorithm on the same experimental platform, using the same set of four files, each one with 1000, 10000, 100000 and 1000000 sequences and ranging k = 2 up to 5. We use a workstation composed by a CPU Intel(R) Core(TM) i7 CPU X 980 3.33GHz with 12 GB RAM memory and one Nvidia Tesla C2050. As results, our experiments over performed 16 times Jellyfish algorithm for k = 2 and 7 times for k = 5. We conclude that the GPU technology is a good approach to reduce the execution time for metagenomics experiments for k up to 5.

Provenance-based Profiling of Swift Parallel and Distributed Scientific Workflows

Maria Luiza Mondelli, Fabrício Vilasbôas, Kary Ocaña, Marta Mattoso, Michael Wilde, Ana Tereza Vasconcelos, Luiz Gadelha

National Laboratory for Scientific Computing, Federal University of Rio de Janeiro, Computation Institute, Argonne National Laboratory/University of Chicago

Abstract

The demand for high-performance computing (HPC) resources has increased in recent time due the complex features of bioinformatics experiments and the biological big data that need to be processed. In general, these experiments execute a set of applications as a flow of activities in which one data is the entry of another activity, suggesting they can be modeled as scientific workflows. Scientific Workflow Management Systems (SWfMS) are used to manage the distribution and parallelism of scientific workflows in HPC environment. Swift is a SWfMS that follows the functional programming paradigm with implicit parallelism. However all benefits provided by Swift, managing provenance scientific data is still an open and challenging problem to be resolved in the next years. Swift creates by default a set of log files containing information of some environment statistics related to the workflow execution. Log files are created as the workflow finishes and contains information e.g., about the workflow execution time or the status of the activity/task execution, which are stored in a relational database. This information is not naturally reported to scientists, but it contains invaluable information about the performance of workflow execution. If scientists could access this provenance, they could be better positioned about their own experiments, for example to determine which data/parameter could generate possible executions errors or if any debugging process can be implemented. For assisting scientists at analyzing the performance of their workflow implementations, we designed a profiler tool called SwiftProfiler. It was implemented in Python and encapsulates a set of SQL queries to the provenance database, to extract resource usage behavior. With SwiftProfiler scientists are able to: (i) calculate execution time of e.g., tasks, activities or total workflow, (ii) report the CPU usage, (iii) trace the provenance of the workflow execution results to each workflow activity, and (iii) present statistical calculation as tables or graphics. As a future work, we will attempt to extend the profiler for processing more sophisticated queries, such as enabling scientists to profile read/write to filesystem behavior.

POTION: an end-to-end pipeline for positive Darwinian selection detection in genome-scale data through phylogenetic comparison of protein-coding genes

Jorge Hongo, Giovanni de Castro, Leandro Cintra, Adhemar Zerlotini, Francisco Lobo

Embrapa Informática Agropecuária

Abstract

Detection of genes evolving under positive Darwinian evolution in genome-scale data is nowadays a prevailing strategy in comparative genomics studies to identify genes potentially involved in adaptation processes. Despite the large number of studies aiming to detect and contextualize such gene sets, there is virtually no software available to perform this task in a general, automatic, large-scale and reliable manner. This certainly occurs due to the computational challenges involved in this task, such as the appropriate modeling of data under analysis, the computation time to perform several of the required steps when dealing with genome-scale data and the highly error-prone nature of the sequence and alignment data structures needed for genome-wide positive selection detection. We present POTION, an open source, modular and end-to-end software for genome-scale detection of positive Darwinian selection in groups of homologous coding sequences. Our software represents a key step towards genome-scale, automated detection of positive selection, from predicted coding sequences and their homology relationships to high-quality groups of positively selected genes. POTION reduces false positives through several sophisticated sequence and group filters based on numeric, phylogenetic, quality and conservation criteria to remove spurious data and through multiple hypothesis corrections, and considerably reduces computation time thanks to a parallelized design. Our software achieved a high classification performance when used to evaluate a curated dataset of Trypanosoma brucei paralogs previously surveyed for positive selection. When used to analyze predicted groups of homologous genes of 19 strains of Mycobacterium tuberculosis as a case study we demonstrated the filters implemented in POTION to remove sources of errors that commonly inflate errors in positive selection detection. A thorough literature review found no other software similar to POTION in terms of customization, scale and automation. To the best of our knowledge, POTION is the first tool to allow users to construct and check hypotheses regarding the occurrence of site-based evidence of positive selection in non-curated, genome-scale data within a feasible time frame and with no human intervention after initial configuration. Our software was recently published in BMC Genomics (http://www.biomedcentral.com/1471-2164/16/567) and is available at `http://www.lmb.cnptia.embrapa.br/share/POTION/`. This work was supported by Embrapa (Brazilian Agricultural Research Corporation), LMB (Laboratório Multiusuário de Bioinformática) and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) [grant number 485279/2011-8].

Identification of snoRNAs using EDeN

João Victor de Araujo Oliveira, Fabrizio Costa, Maria Emília M. T. Walter, Rolf Backofen

University of Brasilia, Albert-Ludwigs University of Freiburg

Abstract

Machine learning methods have been widely used on identification and classification of different families of non-coding RNAs, e.g., snoReport. Many of these methods are based on supervised learning where some previous known attributes, called features, are extracted from a sequence, and then used in a classifier. Instead of using known features from a sequence (difficult to find in general) to identify ncRNAs, a recent approach in machine learning is described as follows. Given a region of interest of a sequence, the objective is to generate a sparse vector that can be used as micro-features in a specific machine learning algorithm, or it can be used to create powerful features on previous methods. One method that uses this approach is EDeN (Explicit Decomposition with Neighbourhoods). EDeN is a decompositional graph kernel based on Neighborhood Subgraph Pairwise Distance Kernel (NSPDK), that transforms one graph in a sparse vector decomposing it in all pairs of neighborhood subgraphs of small radius at increasing distances. In this work, we present a new method based on EDeN to identify the two main classes of snoRNAs, C/D box and H/ACA box snoRNAs: transforming specific secondary structure regions of snoRNA in a graph representation used to build sparse vectors that can be used on different machine learning algorithms, e.g., stochastic gradient descent (SGD). Preliminary results on C/D box snoRNAs classifier showed F-score of 89.5%, Average Precision of 93%, and AUC of 97.5%. Furthermore, this new method discarded 13% less positive samples in the pre-processing phase, when compared to snoReport 2.0, allowing to discover a diversity of snoRNAs quite different to canonical snoRNAs.

Visualizing Probabilistic Suffix Tree

Fábio Sano, André Yoshiaki Kashiwabara

Universidade Tecnológica Federal do Paraná - UTFPR

Abstract

There are many tools that use probabilistic models to the task of biological sequences analysis. An example of its application is ab initio gene prediction, which allows us to predict the location of protein-coding genes in a genome sequence. Some models are not easily interpreted due to the huge amount of parameters, requiring methods to assist our understanding. The data visualization is a very common technique in computer science, and the application of this technique in probabilistic models allows a large volume of parameters to be presented in an organized and grouped form, which facilitates the analysis of the probabilistic models. However, the application of the data visualization, when static, may not be enough to a good comprehension of some models, since they can not expose all the model data in a single view. We seek to develop a novel solution to the problem of visualization of probabilistic models, in particular to the VLMC model (variable length Markov chain). ToPS (Toolkit of Probabilistic Model of Sequence) is a framework to manipulate probabilistic model and ToPS Visualization is an interactive graphical tool to visualize VLMC. This model is represented by using a probabilistic suffix tree which structure can show us important sequence features. In this poster we present two ways to visualize the VLMC: (i) radial tree; (ii) triangular tree. The user can also change the location of any node by dragging it to different locations of the screen. The tool also implements the function of zoom in and zoom out, and allows searches for specific branches or nodes through the sequence of the desired context. We tested the visualization using a VLMC trained with 1,000 random selected sequences of CpG islands from the human genome hg18. The interactive visualization of the model simplified the identification of important patterns of the CpG Island, allowing us to observe all their probability distributions, length of branches, and important sequences. Future work consist in implement novel visualizations for other probabilistic models such as hidden Markov models, and Inhomogeneous Markov Models.

Medical Data Access Accountability in EHR Systems, A Practical Perspective

Paulo Batista, Daniel Grunwell, Tony Sahama, Sérgio Campos

UFMG, Queensland University of Technology

Abstract

The world has experienced a large increase in the amount of available data. Therefore, it requires better and more specialized tools for data storage and retrieval and information privacy. Recently Electronic Health Record (EHR) Systems have emerged to fulfill this need in health systems. They play an important role in medicine by granting access to information that can be used in medical diagnosis. Traditional systems have a focus on the storage and retrieval of this information, usually leaving issues related to privacy in the background. Doctors and patients may have different objectives when using an EHR system: patients try to restrict sensible information in their medical records to avoid misuse information while doctors want to see as much information as possible to ensure a correct diagnosis. One solution to this dilemma is the Accountable e-Health model, an access protocol model based in the Information Accountability Protocol. In this model patients are warned when doctors access their restricted data. They also enable a non-restrictive access for authenticated doctors. In this work we use FluxMED, an EHR system, and augment it with aspects of the Information Accountability Protocol to address these issues. The Implementation of the Information Accountability Framework (IAF) in FluxMED provides ways for both patients and physicians to have their privacy and access needs achieved. Issues related to storage and data security are secured by FluxMED, which contains mechanisms to ensure security and data integrity. The effort required to develop a platform for the management of medical information is mitigated by the FluxMED's workflow-based architecture: the system is flexible enough to allow the type and amount of information being altered without the need to change in your source code.

A multi-agent system performing RNA-Seq analysis

Julien Jourde, Taina Raiol, Maria Emilia Machado Telles Walter, Marcelo de Macedo Brigido

Universidade de Brasilia, Fundação Oswaldo Cruz

Abstract

A multi-agent system (MAS) contains several intelligent agents that interact and work together or competitively. In a cooperative mode, each agent is working asynchronously with regard to the others. They operate autonomously using an infrastructure, called platform, that allows communications and interaction protocols between them. We attempt to propose such a system to perform various analysis on high throughput RNA sequencing results (RNA-Seq). This type of study has become progressively more used those past few years which led to a multiplication of pipelines, integrating different softwares (TopHat or Segemehl.x for the mappers, EdgeR, DESeq and/or DESeq2 for the statistical analysis, for example), and an exponential increase of the amount of data. The principal aim of the MAS is not to substitute itself to the bioinformatician but to help him extract new insights from the results obtained. We already proposed a representation of the early and later requirements of our system using the Tropos methodology. A prototype able to perform the early first step of an RNA-Seq analysis, meaning checking the quality of the reads using the FastQC tool, is already running. To calibrate and test our system, we are only using existing biological data of human T cells provided by other projects from our lab or published data. Two different pipelines are used on two different sets of data and we intend to compare the results obtained without the MAS with the one obtained through it using a simple graphical online interface. A subsequent result of these comparisons would be new informations on the RNA isoforms discovered by these different analysis.

Modelling Data-intensive Metagenomics Experiments Using Scientific Workflows

Silvia Benza, Kary Ocaña, Vitor Silva, Daniel De Oliveira, Marta Mattoso

COPPE, Federal University of Rio de Janeiro, Bioinformatics Laboratory, National Laboratory for Scientific Computing, Computing Institute, Fluminense Federal University

Abstract

Metagenomics analyses study the genetic material of environmental samples that are directly gathered from a single or a collection of genomes without the need to cultivate them in wet labs. Metagenomic analyses are supported by computational analyses. Designing metagenomic analyses is a laborious task since it involves the exploration of several simulation flows, in which each application within the flow consumes and generates a large amount of data. The most representative programs used in the identification of protein candidates in metagenomic genes are Prodigal, MetaGeneMark, FragGene, Glimmer, GeneMarkS, and BLAST. These programs are usually classified into two main types: the ab initio heuristic algorithms and the programs based on genome comparisons. One efficient way to validate the novel identification of potential proteins is to use both types of algorithms. Then, the findings are more likely to be consistent since they are validated by both approaches. However, exploring all these algorithms within the same experiment is not simple, as it involves at the same time the manipulation of several programs in a coherent execution strategy. Metagenomics analyses can by modeled as scientific workflows and managed by Scientific Workflow Management Systems (SWfMS). In this work, we propose a scientific workflow for finding genes based on the aforementioned applications, thus exploring different algorithms. The proposed workflow, named SciMG, was implemented in the SWfMS SciCumulus benefiting from its cloud-based parallelism and its steering mechanisms. Steering mechanisms allow for adding the Human In the Loop (HIL), i.e. scientists can interact with the running experiment. Although the current version of SciMG is modeled to execute with minimal human interaction, in some activities it is required the specialist knowledge to advance in the experiment. With this requirements established, we aim at providing an infrastructure capable of offering a solution that supports metagenomics analyses and includes the HIL. Allowing the decision making to be dynamic and entirely up to the bioinformatician, wherein the SWfMS is used in the background hiding the well-know difficulties found in metagenomic analyses executions.

Labcontrol: A software for bacterial information management

Mariana Parise, Flávia Rocha, Roselane Gonçalves, Doglas Parise, Elma Leite, Anne Pinto, Vasco Azevedo

Universidade Federal de Minas Gerais, Universidade Federal de Pernambuco

Abstract

The transition of traditional in vivo discovery methods to in silico techniques changed the study of life sciences. Currently, "Omics" rise and new techniques development lead to genetic data exponential growth which generates a management bottleneck to such data. Without efficient management of these data, it may become obsolete or even be lost. In this context, a demand for "omics" data computerization appears in order to present such data in a clear and concise manner. LabControl software have been designed according to a study about the researcher's needs in the Laboratory of Cellular and Molecular Genetics (LGCM) located in UFMG. Eliciting requirements step were performed through interviews and investigation of user's routine, which results in a more reliable data model, this approach provides better assistance from software to the researchers' real needs. LabControl software aims to assist bacterial collection management and organization, by normalizing and integrating different organisms data, as well as administrate datasets generated by sequencing pipelines and in vitro or in vivo techniques applied to strains. A sequencing pipeline comprehends data related to sequencing, assembling, annotating, submitting and publishing for one or more strains. For each pipeline step there is stored information regarding status, researchers involved and specific description. This software has been developed as a minimalist web system, focused on user and improved usability based on the software development methodology extreme programing. The technologies used in the development of LabControl are the Java programming language in NetBeans IDE integrated with Spring and Hibernate frameworks, using the PostgreSQL database. LabControl software allows users to faster strain information management through an easy use system, which reduces the impact of solution migration from spreadsheets to a web system. This migration may promote a significant gain in researchers' daily productivity using data organized, updated and persistent of all studied strains.

Applying Data Mining Techniques to Identify Frequent Phylogenetic Trees using Scientific Workflows

Kary Ocaña, Lygia Costa, Daniel de Oliveira, Vitor Silva, Marta Mattoso

National Laboratory for Scientific Computing Fluminense Federal University, Federal University of Rio de Janeiro

Abstract

The volume of available genomic data is growing in a fast pace due to the recent improvements in biological sequencing and High Performance Computing (HPC) technologies. This fact can be seen as a challenge for several large-scale bioinformatics analyses that need to extract and infer knowledge on these huge amounts of data. Phylogenetic reconstruction experiments produce trees and statistic calculation used for inferring the evolutionary history and relationships of species. One important task in phylogenetic experiments is to identify frequent generated subtrees because they represent shared species among several genes. This information can be used to aid drug development, especially when these genes are involved in crucial metabolic cycles. The generation of trees and its analysis is compute and data-intensive since it processes a huge amount of non-structured data. Although engines such as Scientific Workflow Management System (SWfMSs) helps executing and managing experiments they fail to provide support for mining and analyzing this data. Identifying frequent trees from the produced phylogenetic results is far from trivial, as it can involve the exploration and comparison of thousands of final/intermediary files. Usually, scientists attempt to infer this knowledge manually but this process is laborious and error prone. Data mining techniques can aid the scientists' analysis by extracting patterns and hidden information. For the present study, we aim at mining the huge amount of phylogenetic trees (i.e. several different trees) and by identify common patterns, i.e. frequent sub-tress on the produced trees. We added to the phylogenetic analysis workflow named SciPhy, activities for mining the phylogenetic data. SciPhy was modeled using the cloud-based SWfMS SciCumulus, benefiting from its features as parallelism, fault-tolerance and provenance management. By using SciCumulus, all produced trees are stored in the provenance database (and can be queried). Then, scientists are able to e.g., trace the taxonomic lineage of one taxon inside the same or several trees only by querying the database (e.g. determining the child nodes from a parent inside the same tree, searching for which subtrees are shared by different trees or extracting all trees/subtrees with the highest bootstrap value) to extract the input data for the data mining algorithms. By using the proposed approach in an experiment, we could identify 2 frequent subtrees in a set of 100 phylogenetic trees of 20 species. We believe that the contribution of this article can assist specialists in analyzing the hundreds or thousands phylogenetic trees in a more efficient manner.

CeTICSdb: Integrated analysis platform for high-throughput -omics data and mathematical modeling of molecular signaling networks

Milton Y Nishiyama-Junior, Marcelo da Silva Reis, Daniel F Silva, Inacio L M Junqueira-de-Azevedo, Julia P C da Cunha, Junior Barreira, Leo K Iwai, Solange M T Serrano, Hugo A Armelin

Instituto Butantan, Universidade de São Paulo

Abstract

The Center of Toxins, Immune-response and Cell Signaling (CeTICS) includes researchers from several areas to studies biochemical, molecular, and cellular mechanisms of toxins that have therapeutic potential, aiming to understanding the behavior of biological systems based on analysis of high-throughput data and signaling networks. The research and analyses developed in CeTICS are intrinsically interdisciplinary, which is coupled to the heterogeneous, high-throughput data from genomics, transcriptomics and proteomics, and implies the necessity of data organization and integration to carry out scientific investigations. High throughput data and biological knowledge has increased rapidly and this information has to be crossed with signaling diagrams and biological annotation, in order to mine meaningful results. The CeTICSdb aims to provide a dynamic, user-friendly management system that fully support researcher management (e.g., multiples privileges and roles), data management (e.g., data submission attached to its semantics, automatic preprocessing, visualization and sharing), automate customized real-time analyses and simulations and apply multivariate methods to the integration and comparison of multiple datasets on same and different –omics technologies. The CeTICSdb has been developed in an integrated Django-based platform, integrating Galaxy and GBrowse applications for interdisciplinary research that allows data management, quantitative and qualitative –omics analysis and mathematical modeling of biochemical reactions (e.g., metabolic pathways, molecular signaling networks), based on semantic data models in a relational database. It is currently running in a high performance server, which processes all user requests through a cluster queuing system. However, none of the existing tools completely address the needs for combining and integrating heterogeneous data. To validate the platform, we integrated transcripts or protein expression profile with Metabolic Pathways to: i) estimate the metabolic activity between different conditions or treatments; ii) define and compare the functional activity for the metabolic pathways in each condition. Finally, our mid-term objective is to make the CeTICSdb platform available as a dry lab to the scientific community.

AQUAtigs: an interactive web-tool for scaffold contigs and complete bacterial assemblies

Felipe Pereira, Carlos A. G. Leal, Henrique César Pereira Figueiredo

Aquacen/UFMG Universidade Federal de Minas Gerais

Abstract

The advances on Next-Generation Sequencing (NGS) benchtop machines resulted in the generation of an uncountable number of genomes. A high number of these genomes are of bacterial origin and were deposited as drafts in public access banks. While obtaining drafts can be done by automatic processes in days or hours, finishing a complete genome can take months and is an expensive process. The usual short reads assembly processes produces a variety number of contigs that can be correlated to complex genetic artifacts like ribosomal RNA, operons, transposons, phages, and TANDEM repeats present on the genome. Drafts of bacterial genomes usually contains at least 95% of the nucleotide sequence corresponding to the genomic DNA, but not ordered as an unique chromosome. To mitigate this plentiful number of contigs, a posterior scaffolding process is performed, using a closely related organism or information from other technologies (e.g., optical maps). After that, a gap filling process is conduced to reduce or zero out the unassembled regions. At this work, we introduce AQUAtigs, an interactive web-tool to order contigs on a single scaffold and perform a gap filling step. AQUAtigs is user-friendly software that, by inputing raw data and an assembler output, it orders the contigs (based on Blast+ analysis) and fills the remaining gaps (evaluating overlaps and mapping reads on contig flanks). AQUAtigs is designed to work in internet browsers and consequently is an useful tool to naïve computer users. This tool was tested with three bacterial genomes, Francisella noatunensis subsp. orientalis, Corynebacterium pseudotuberculosis and Streptococcus agalactiae and was shown to be an efficient strategy to help finish bacterial genomes. The final chromosomes assembled by AQUAtigs present 3, 13 and 23 gaps, respectively, in hands-on time mean of 50 minutes.

IN SITU TRANSCRIPTIONAL PROFILE TO DISCRIMINATE RESPIRATORY INFECTION CAUSED BY VIRUS AND/OR BACTERIA IN CHILDREN WITH ACUTE RESPIRATORY INFECTION.

Kiyoshi Fukutani, Ricardo Khouri, Tim Dierckx, Johan Weyenbergh, Camila Oliveira

FIOCRUZ, KULEUVEN

Abstract

Acute respiratory infections (ARI) caused by microbes affect the respiratory system and lead to high mortality rates. In recent years there has been increased interest in the development of rapid and accurate diagnostic tests for detection of respiratory pathogens and for the identification of biomarkers associated with eitherdistinguish viral and bacterial infections. The present study usesthe transcriptional approach and aims at identifying such biomarkers. This prospective cohort study employednasopharyngeal aspirate samples of children with acute respiratory infected attending the Emergency Room of Federal University of Bahia Hospital. A custom-designed nCounter probeset containing viral and bacterial targets was tested to identifythe infection type and the accompanying human immune response. We detected the presence of viral and bacterial transcripts in the transcriptomic signatureof75 patients. Of these, 27 patients presented bacterial infection only (BI), 20 patients presence bacterial and viral coinfetion(CI) and4 patients presented viral infection only (VI). In these same patients, we compared the expression of immune response genes against 3 healthy control samples. We found 87 genesdifferentially expressed in patients withBI,108 genes in patients with CI and 75 genes in patients withVI. Enrichment pathway analysis showed the presence of different pathways (LXR/RXR Pathway, Role of Pattern recognition of bacteria and viruses and IRF activation pathway) depending on the type of infection. The activation of different pathway had fewnon common genes (3 genes for BI, 26 for CI and 4 genes for VI). We used these genes in a canonical discriminate analysis and we were able to distinguish VI, BI and CI with 99% AUC forBI and CI and 100% AUC for VI. The standardized scoring coefficients suggests a key role forLILRA1, LILRA3, LILRA5, LILRB3,LILRB2, CCRL2, CCL23, IL1A, IL6,TNFAIP3, IDO1, SERPING1, NLRP3, ICAM1, FCGR2C, TLR2 and PLAUR in discriminating these infections in situ. nCountertranscriptomicscan, in a uniquecodeset,detect pathogens and theimmune response in nasopharyngeal aspiratesin situ.This work was supported by FAPESB and CNPq.

Evolutionary aspects of gene expression during Drosophila melanogaster spermatogenesis

Júlia Raíces, Maria Vibranovski

Universaidade de São Paulo

Abstract

Genes that appeared recently in the evolutionary history of a taxonomic group are considered new genes. They can arise by exon recombination, transposons, lateral gene transfer, gene fission/fusion, DNA based duplications, retroposition, de novo origination or combinations of those mechanisms. Although some of those new genes are functional, most of them become pseudogenes. However, when they happen to be functional, new genes can quickly convert to essential genes or bear an important function at different phases of the development of individuals of different species. In this way, new genes expressed during spermatogenesis – the system of male gamete development - are probably related to fertility, mobility, form and function of the sperm cells. And, therefore, must be more expressed during the late phases of the gamete development, bearing in mind that expression relates to functionality. Hence, this project aims to test the hypothesis that there is a relation between a gene's age and its expression during spermatogenesis. To test this hypothesis, we searched for a correlation among genes evolutionary ages and it's differential expressions in spermatogenesis phases in Drosophila. According to this hypothesis, it's expected that new genes are more frequently expressed during post-meiosis, the latest phase of germline development. To answer those questions, bioinformatics and computational biology tools were used and statistical methods were applied to correlate already available data of gene age and gene expression during spermatogenesis phases in Drosophil. Our results in Drosophila have shown that the proportion of new genes expressed in late expermatogenesis (meiosis and post-meiosis) is significantly higher then in the beginning of the processes (mitosis). Also, we found that the expression level of new genes is higher than the expression of old genes during meiosis and post-meiosis, and the opposite pattern occurs during mitosis. These results implicates that new genes have an important role during the late spermatogenesis, which could be related to sperm fertility and speciation process.

Differential expression and functional associations of Hox genes

Rodolpho Lima, Christiane Nishibe, Tainá Raiol, Nalvo Almeida

Federal University of Mato Grosso do Sul, Oswaldo Cruz Foundation

Abstract

Hox genes are evolutionarily highly conserved among different species and organized into genomic clusters, normally located in different chromosomes. They code for proteins that act as regulatory transcription factors during embryogenesis, activating or repressing their downstream targets. Recent works have reported that Hox genes also play important role in the development of cancer. They are dysregulated during development of several solid tumors, including colon, breast, prostate, and kidney cancers, with different patterns of changes, depending on the type, the location or even the stage. Besides, non-coding RNAs within the Hox clusters regulate the expression of Hox transcription factors. Investigating regulatory mechanisms of Hox genes, their co-related and co-located ncRNAs, and also selected genes of specific pathways down or up-regulated following differential expression of Hox genes may provide researchers a better understanding of potential new targets for cancer therapeutics. The purpose of this work is two-fold. First we present a specific pipeline for differential gene and transcript expression analysis of RNA-seq experiments in Hox clusters, to describe in details differential expression of Hox genes and all types of ncRNAs across them for further studies. We intend to exploit the close proximity of the genes and provide a high-quality genome browser map of transcripts of these loci, small enough to be examined by eyeballing. Second we implemented, using Javascript and APIs, an interactive web tool showing a MA Plot, containing all Hox genes, the ncRNAs co-located in Hox clusters, all co-expressed by them and finally all genes of interest given by the user (whether or not genes in these last two sets are in the clusters). All genes in the interface are are clickable, linked to databases of protein interactions and gene ontology, like STRING and COG, respectively, allowing further GO functional enrichment. Tests using our tools have been done with cancer-sequencing data. We also tested them with another kind of data sets, like the differential expression of ncRNAs located in Hox clusters that are related to Huntington's Disease. Although the initial purpose is to set up a platform to study Hox and their co-related genes, the interactive MA Plot can be used to any kind of differential expression data set, given a full list of genes of interest, since all of them will be shown in the graphical tool, wherever they are positioned on the chromosomes.

Quantification of Whole Transcriptomes by Ion Torrent

Vitor Coelho, Michael Sammeth

Universidade Federal do Rio de Janeiro

Abstract

Transcriptomics aims to characterize the molecular phenotype by the type and abundance of each expressed gene. So far, RNA-Seq on the most popular platforms (by Illumina) focuses exclusively on one biotype of RNA, experimentally restricted by the length selection of the (fragmented) RNA. However, whole-transcriptome analysis with total RNA molecules sequencing (total RNA-Seq) enables to study a widespread range of gene expression and transcripts, e.g. mRNA, rRNA, tRNA and other non-coding RNAs. Recently, the Ion TorrentTM (IT) sequencing platforms have been introduced, which in contrast to the Illumina standard are producing the reads with variable length. Since the read length can reach up to > 300 nucleotides (nt), the read length can reflect well the size of the sequenced molecule or fragment. With a theoretical capacity of 100 million reads or more, the most recent IT Proton platform is also able to generate the amount of reads required for quantitative RNA-Seq. Thus, for the first time we have the opportunity to quantitatively analyze a broad spectrum of RNAs with variable length in a single experiment, since miRNAs to mRNAs. This work aims to study the development of new bioinformatics techniques for quantitative analysis of transcriptomes using the IT platform. Our goal is to extend bioinformatics methodologies and statistics models for comparing gene expression considering the variability of read lengths in the experiment as an additional parameter. Beyond novel challenges in dealing with the types and frequencies of errors introduced by the IT sequencing chemistry, that have been shown in multiple preliminary works, our first studies demonstrate that sequencing whole transcriptomes on the IT imposes several challenges caused by the varying attributes between different RNA biotypes, principally caused by their markedly different molecule lengths. For instance, reads from short RNAs are delimited in their length by the size of the original molecule (< 30 nt for the smallest RNAs). Difficulties in mapping such short reads can directly introduce biases in the quantitative estimates of the corresponding genes, and also our results show that the standard RPKM-measure can vary substantially between replicates of different biotypes when assessed by the IT. In this light, we are revising common metrics like RPKM (Reads Per Million per kilobase mapped reads), FPKM (Fragments Per Million per kilobase mapped reads) and TPM (Transcripts per Million) in their use with RNA-Seq experiments produced by the IT platform.

Integration of two pipelines for annotation and detection of miRNAs in platelet concentrates stored at blood bank

Jersey Maués, Caroline Moreira-Nunes, Thaís Pontes, Letícia Lamarão, Rommel Burbano

Universidade Federal do Pará

Abstract

The detection of miRNAs in platelet concentrates (PC) can control the discharge of bags important CP using the miRNA expression profile suffering variations and modulate the translation of mRNA-target related to platelet storage lesion. We adopted the sequencing procedure with the Solexa Genome Analyzer IIx six pools of PC used for extraction miRNAs that were obtained from six different days of storage at 22 $\pm$ 2°C. The miRNAs were extracted from 16 fragments of different PC pipes, making six pools referring homogeneous rates of 16 bags of 80 donors gathered for six days (1st, 2nd, 3rd, 4th, 5th and 7th) experiment in HEMOPA Foundation. Our data were analyzed with two templates pipelines. The first analysis treated the raw reads 17-35-nt that were generated in the sequencing in fastq format and were cleaned with CutAdapt tool. The high quality reads were aligned and mapped to the human genome with the STAR tool. In the second stage of the pipeline was used to miRSeq tool. We treat the raw data and readPro readMap to give a good cleaning quality sequencing, mapping and annotation mature miRNAs from known precursors. The raw data platelet sequences are available from GEO (GSE61856). The total of precursors reached a total of 1,273,288 and 5,228,808 mature equivalent to 36% with the greatest amount of expressed mature miRNAs. The largest amount of miRNAs were detected in the first collection period PC, representing 807 mature miRNAs and 612 precursors, equivalent to 36% of the readings detected with readPro. With readMap were generated an amount of isomiRs 3p and 5p. The expression profile of miRNAs identified in the fifth period PC, showed that eight miRNAs were lost: miR-127-3p, miR-103a-3p, miR-221-3p, miR-151a-3p, 5p, miR-99b, let-7g-5p, miR-4433b-5p, miR-28-5p. Therefore, the increase of miRNAs: hsa-miR-127-3p, hsa-miR-26a-5p, hsa-miR-22-3p, hsa-miR-423-5p, let-7g-5p in the last period is an indication that the PC bags were still physiologically normal platelets.

Functional annotation of families of miRNAs expressed in progressively extended periods platelet concentrate (PC)

Jersey Maués, Caroline Moreira-Nunes, Thais Pontes, Letícia Lamarão, Rommel Burbano

Universidade Federal do Pará

Abstract

In platelets more 492 different functional mature miRNAs were detected, characterized by a small highly expressed and represented by families number, such as members of the let-7, the most abundant being a percentage of 48% of the content of miRNAs platelet consistent with the role of let-7 in cell differentiation and impact on differentiation of megakaryocytes. The Solexa sequencing by Genome Analyzer IIx platform for samples from fresh human platelets PC healthy patients were generated slightly more than 16 million reads. We obtained precursors mapped around total of 1,273,288 and 5,228,808 to mature miRNAs, equivalent to 36% with the greatest amount of expressed mature miRNAs. From downloaded from miRBase V20 files that have been manipulated to note the families of miRNAs precursors of the human species (hsa), this procedure was used for annotation the miRNAs PC. Our approach with high-throughput sequencing (HTS) of platelet (miRnome) reveals the diversity and relative abundance of miRNAs. The most abundant family was mir-486 which represented 60% of miRNAs platelet content, as well as other families of miRNAs that were highly representative in PC. In addition to the mir-486 family, the 10 families of the expressed miRNAs were: mir-191, let-7i, mir-92a, mir-181a, let-7a, let-7b, mir-320, mir-127 and mir -26a. Functional annotation with DAVID V6.7 (Bioinformatics tool) targets the mature miRNAs revealed that they modulate more 60 routes and cellular mechanisms of platelet activation and apoptosis. In all periods of the PC have been identified highly expressed miRNAs miR-486-5p, miR-191-5p, let-7i-5p, miR-92a-3p, miR-181a-5p, let-7a-5p, let- 7b-5p and miR-320a, which can act as endogenous inhibitors for impairment of PC and assist in the validation and disposal bags stored for extended periods.

CAUSES AND EFFECTS OF ALLELE-SPECIFIC EXPRESSION

Cibele Masotti, A Buil, A Brown, A Vinuela, M Davies, HF Zheng, JB Richards, KS Small, R Durbin, TD Spector, ET Dermitzakis

Instituto de Ensino e Pesquisa do Hospital Sirio-Libanes, University of Geneva, Wellcome Trust Sanger Institute, King's College London, King's College London, McGill University

Abstract

There are several tools that can predict the impact of a genetic variant on the structure and function of a protein, but the penetrance and expressivity of the trait associated to these variants remain unpredictable. It has been demonstrated that regulatory variants can contribute to phenotypic variation by modulating the expression of a linked coding variant; therefore, the expression levels of functional/pathogenic alleles could be predictive of phenotype outcome or severity. However, patterns of allele-specific expression (ASE) can vary among tissues and most mechanisms by which levels of allele expression interfere in phenotypes need to be explained. Our working hypothesis is that ASE has an effect on the expression of downstream or related trans genes. By using transcriptome (RNA-seq) data from different tissues (whole blood, skin, fat, and lymphoblastic cell lines) of 856 individuals from the EuroBATS study we measured allelic expression of SNPs and then correlated allelic ratios with genome-wide expression. Through this strategy we directly measured ASE effects on expression of 10,000 distal transcripts per tissue, allowing the identification of novel disease candidate genes related to differential expression of GWAS-variants. A striking example is the PRNP (prion protein gene) GWAS-variant rs1799990, previously related to neurodegenerative diseases: the strongest trans-association was with OLFML3 (p-value=5.8e-06, 30% FDR), already known to co-express with PRNP in vesicles from the cerebrospinal fluid. We also investigated whether trans-effects of allelic ratios were maintained across different tissues. In general, we observed that trans-associations are tissue-dependent. Regarding the PRNP-OLFML3 example, this association was specifically observed in skin, a tissue that shares embryological origin with brain. Another way to interpret an association between allelic ratios and trans-gene expression is to consider that the trans-gene can regulate the gene in which we observed ASE, i.e., instead of being a "downstream" gene, in fact operates "upstream" and modulates the rates of allelic expression. In this context, we have also tested a model where allelic ratios are explained by an interaction between cis-eQTL genotypes and trans-gene expression. Through this model, we were able to identify 235, 265, 28, and 300 putative cis X trans interactions (1% FDR) in LCLs, fat, skin, and blood, respectively.

Schistosoma mansoni lincRNAs mining from NGS data

Elton Vasconcelos, Bruno Souza, David Pires, Murilo Amaral, Sergio Verjovski-Almeida

University of Sao Paulo, Instituto Butantan

Abstract

Schistosomes are flatworm flukes that cause the infectious and parasitic disease known as schistosomiasis. They are blood parasites widely distributed around the world with a high importance for Public Health and studies in the field of Molecular Parasitology. The genome of Schistosoma mansoni (etiologic agent in Brazil), which was completely sequenced in 2009, has a size of 363 Mb and over eleven thousand genes were already mapped. These static data pose the even greater challenge of understanding the molecular dynamics responsible for the peculiar features of the parasite biology. Post-transcriptional control of gene expression events, such as modulation of mRNA alternative splicing and silencing by RNAi, as well as epigenetic events such as histone modifications, are present in Schistosoma and, as in other higher eukaryotes, it is believed that a variety of regulatory non-coding RNAs (ncRNAs) mediate such reactions. Next Generation Sequencing (NGS) strategies, like RNA-Seq, have revealed the presence of a wide variety of ncRNAs in the genomes of several organisms. These assays are able to generate a large volume of data that are prone to computational analyses that seek to reveal key biological information about gene expression and its regulation. We are focused on conducting in silico experiments aimed at assessing the non-protein coding transcriptome content, with emphasis on long intergenic ncRNAs (lincRNAs), hidden in the large amount of information obtained through RNA-Seq in Schistosoma. We idealized and built an in silico pipeline that starts with any transcriptome assembly input, maps it to the genome and is capable of identifying a reasonable and reliable set of putative spliced lincRNAs represented by the assembled contig's sequences. RT-qPCR assays have confirmed that the candidates tested so far exhibit lncRNA traits, such as a high cycle threshold (CT > 29), i.e. exhibit low expression, comparable with the lowexpression (CT = 30) or the medium-expression (CT = 25) protein-coding genes. Further studies, directed to a systems biology holistic overview that merge non-coding content with what is encoded into proteins and their interrelated regulatory pathways, are going to be addressed on the attempt of generating concrete hypotheses on Schistosoma-specific gene regulation mechanisms that might pinpoint elucidative target-driven molecular assays.

Searching for variations of pharmacological receptors of Praziquantel and potential targets for new drugs against Schistosoma mansoni

Jéssica Hickson, Fabiano Pais

Faculdade Infórium de Tecnologia

Abstract

The Praziquantel (PZQ) is currently the only recommended drug for the treatment of schistosomiasis infection. The drug has low activity against younger forms of the parasite and, in specific cases, has been appointed as a ineffective medicine due to the emergence of strains that have developed resistance against mechanism of action. Some calcium channel Cav type were determined as pharmacological receptors of PZQ. However, several studies are still being conducted in orther to confirm the association between the drug and targets. In this study, we evaluated the nucleotide sequence conservation of the genes related to PZQ activity, as well as genes that could be related to the drug and/or could be pharmacological receptors of new drugs for the treatment of schistosomiasis. This research is consistent with the fact that sequence conservation is a key requirement for a promising pharmaceutical target since mutations can alter drug interaction with the organism. Eleven RNA-seq libraries of S.mansoni were compared with the reference parasite genome, allowing the identification of non-synonymous variants in SmCav1B gene that can be related to PZQ activity and the SmCavβ gene. Fourteen non synonymous mutations were found in SmCav1B distributed in eight libraries. Schistosomulae were the ones who had this SNVs in SmCav1B gene compared with the other stages of the parasite. Fot the SmCavβ genem we found one MNV within in three libraries. Additional genes (SmCav1A, SmCav2A, SmCav2B, SmCavβvar) showed no synonymous mutations. Regarding the conservation in sequence level, for these specific genes, the lower number of possible mutations could be related to preservation of drug activity. The SmTGR gene, regarded as a more promising target for new drugs, also showed no sequence variations, which reinforces its potential as a therapeutic target. Supported by: Faculdade Infórium de Tecnologia

Bacterial thermostable biomass-degrading enzymes in metagenomic and metatranscriptomic data of São Paulo Zoological Park composting

Roberta Pereira, Luciana Antunes, Aline da Silva, João Carlos Setubal

Universidade de São Paulo

Abstract

The economic interest in thermophilic microorganisms is increasing due to their ability to tolerate extreme conditions during industrial processes. We study a composting process in which the temperature can reach temperatures as high as 80 ^{o}C, and therefore it is a promising source of these thermostable enzymes. Previous work from our group has shown that the composition of the composting microbiota at the São Paulo Zoo Park is highly diverse and its structure varies throughout the process. These observations were based on shotgun metagenomics, high-throughput sequencing of 16S rRNA gene amplicon and RNA-seq of time series samples collected during composting. Analyses of this data also revealed that over 30% of metagenomic sequences from composting samples are derived from bacteria with unsequenced genomes. In the present work, we performed a prospection for genes encoding enzymes responsible for the degradation of biomass using the metagenomic and metatranscriptomic sequence datasets. Our analyses showed that the four categories of enzymes (hemicellulases, cellulases, ligninases and pectinases) necessary for lignocellulose degradation are present in the dataset. Transcripts for these enzymes were detected with an apparently similar expression profile, being higher after 30 days of composting. Furthermore, microorganisms potentially responsible for the expression of such enzymes were identified using program myTaxa. The main genera found were Bacillus, Geobacillus, Paenibacillus and Clostridium. Transcripts encoding xylose isomerases were also identified, mainly on the sample collected after 64 days of composting, and the microorganisms responsible for their expression appears to belong to Thermobacillus, Geobacillus and Clostridium genera. The sequences we have identified will be selected for bacterial recombinant expression and biochemical characterization. This study has the potential for the discovery of novel biotechnologically relevant biomass-degrading microorganisms, as they are potential sources of thermostable enzymes.

Allele specific expression analysis in bovine muscle tissue

Marcela M. Souza, Fabiana B. Mokry, Polyana C. Tizioto, Priscila S. N. Oliveira, Adriana Somavilla, Aline S. M. Cesar, Daniela Moré, Gerson Mourão, Wellison J. S. Diniz, Maurício A. Mudadu, Simone C. M. Niciura, Luiz L. Coutinho, Adhemar Zerlotini, Luciana C. Regitano,

Federal University of São Carlos, UNESP Universidade Estadual Paulista, University of São Paulo, Embrapa Pecuária Sudeste, Embrapa Informática Agropecuária

Abstract

Imprinted genes have been target of many studies, mainly in human and mouse, and lately in bovines due to the interest of understanding the epigenetic mechanisms underlying important meat quality phenotypes and the possibility of applying it in animal breeding programs in the future. Genomic DNA from 146 steers was genotyped using the Illumina BovineHD BeadChip in order to identify heterozygous individuals with known allele origin. Total mRNA from muscle was extracted and sequenced by Illumina HiScanSQ. The software ALEA was used to create a diploid reference genome for each individual, in which haplotype regions were reconstructed from the individual haplotypes. ALEA also invokes BWA to map short sequencing reads to the in silico genome constructed and detects reads that are uniquely aligned only to one of the two haploid genomes. In-house software was developed to compute the frequency of reads mapped to each allele and to perform binomial statistical tests in order to identify allele specific expression. From 742,906 SNPs contained in the Illumina BovineHD BeadChip, 419 were assigned to be imprinted based on the following criteria: heterozygous in the individual, homozygous in its sires, at least 20x RNA-Seq coverage, and $p<0.05$ for the binomial statistical test. The Ensembl software VeP (Variant Effect Predictor) was used to determine the effect of these SNPs on genes, transcripts, and protein sequence, as well as regulatory regions. The VeP report together with phenotype information of the steers will be carefully examined in order to elucidate the molecular mechanisms involved with beef quality.

In silico identification and analysis of noncoding RNAs using RNA-seq data from Leishmania donovani.

Patrícia de Cássia Ruy, Ramon de Freitas Santos, Elton José Rosas de Vasconcelos, Peter Myler, Angela Kaysel Cruz

FMRP-USP, SBRI

Abstract

Protozoa parasites of the family Trypanosomatidae and genus Leishmania cause a group of diseases known as leishmaniases. Trypanosomatids exhibit a remarkable variety of biological processes unique among eukaryotes. The molecular peculiarities include polycistronic transcription and mRNA processing that occurs by trans-splicing mechanism. This genetic organization brings the control of gene expression to the post-transcriptional level. RNA stability and translation rates are central aspects of this regulation and depend on transcript cis-elements and trans-acting factors that may include ncRNAs (noncoding RNAs), which are poorly studied in Leishmania parasites. Developmentaly regulating long-ncRNAs have been described in Leishmania. In addition, we have identified a group of ncRNAs emerging from UTRs. This project aims to develop and integrate computational tools to identify and analyze ncRNAs of Leishmania donovani. The RNA-seq libraries were obtained from three different developmental stages: prociclic promastigotes, metaciclics promastigotes and axenic amastigotes. Three independent assays were performed, totalizing 18 libraries constructed with "TrueSeq® Stranded mRNA" kit and sequenced in MiSeq-Illumina platform. These libraries were generated from the cytosolic and nuclear fractions. We obtained from the cytosolic fraction of procyclics 1,263,336 reads, 495,916 from the metacyclics and 875,192 from the amastigotes. From the nuclear fractions of procyclics, metacyclics and amastigotes we obtained 1,672,108, 1,064,026 and 1,396,664, respectively. The adaptors were trimmed using the cutadapt program and the quality/length of reads were verified with FastQC. Bowtie2 was used to map the reads against the L. donovani genome version 6.0 (TriTrypDB). The percentage of mapped reads was satisfactory, varying between libraries (32%-98%). However most of reads (50% of cytosolic and 80% of nuclear fraction) mapped to the ribosomal RNA locus. Despite the low coverage, the sequencing showed enough results for preliminary analysis. HTSeq program highlighted 37 regions with high-density reads and 16 of these were selected for characterization and experimental analysis. The first Northern blotting experiments confirmed the existence of four small transcripts emerging from UTRs. The characterization showed 11 ncRNA candidates with secondary structure similar to already known ncRNAs (RNAcon); 5 ncRNA candidates with already characterized ncRNA domains (RNAspace); 8 ncRNA candidates with non-coding potential (PORTRAIT). The investigation of these ncRNAs in parasites becomes central because they may play a key role in the parasite survival and pathogenesis. The use of data mining techniques will enable to disclose novel and unidentified patterns of ncRNAs in Leishmania.

Characterizing the alternative usage of spliced leader trans-splicing acceptor sites in Trypanosoma cruzi under gamma radiation stress

André Reis, Mainá Bitar, Priscila Grynberg, Helaine Vieira, Willian Prado, Dominik Kaczorowski, Andrea Macedo, Carlos Machado, John Mattick, Glória Franco

Departamento de Bioquímica e Imunologia, UFMG, Embrapa Recursos Genéticos e Biotecnologia, Garvan Institute of Medical Research

Abstract

Trypanosomatids, such as Trypanosoma cruzi, possess intronless genes and perform polycistronic transcription. The maturation of individual mRNAs is accomplished by coupling spliced leader trans-splicing (SLTS) with polyadenylation. In the SLTS mechanism, a 39-nucleotide sequence is attached to the 5′ end of individual cistrons upon recognition of splice acceptor sites, resolving the 5′UTR. Alternative SLTS has been previously observed in related parasites and is associated with regulation of gene expression. In fact, different roles have been speculated for this mechanism, but a deeper investigation is necessary to gather experimental evidence and clarify how specific splice acceptor sites are selected under diverse environmental conditions. It has been demonstrated that T. cruzi is highly resistant to ionizing radiation, sustaining an exposure to 500 Gy of gamma radiation. Under this circumstance, genomic DNA is fragmented, but the karyotype is gradually restored in the next hours, leading to a complete establishment of the chromosomal band pattern in less than 48 hours. The aim of this study is to characterize the regulatory impact of the ionizing radiation stress on the mechanism of SLTS in this parasite. Thus, a time-series RNA-seq experiment was designed to evaluate acceptor site usage in epimastigotes of T. cruzi CL Brener strain not exposed versus 4, 24 and 96 hours after exposure to 500 Gy of gamma radiation. Each time point was represented by three biological replicates and all samples were sequenced using the Illumina platform. The pipeline applied for the calling of splice acceptor sites was mainly comprised of: FastQC (for quality check), Cutadapt (for identifying and trimming the spliced leader sequence), BWA-mem (for mapping to the reference genomes), Python in-house scripts (for the actual calling of splice sites) and R scripts (for statistical analyses). A total of 38,914 different splice sites were identified using this protocol. The sites were assigned to 16,893 different annotated genes, representing an average of 2.3 sites per gene (median of 2), corroborating the existence of alternative trans-splicing in T. cruzi. Subsequently, the main splice site for each gene will be determined and the differential usage of alternative sites on each time point assessed. Finally, after better investigating how the selection of a particular splice site works, we intend to search for possible effects on gene regulation, such as the incorporation of uORFs, use of alternative translation start sites and inclusion/exclusion of signal peptides.

Transcriptional changes due to Wolbachia infection in the mosquito Aedes fluviatilis reveal evidence of residual host manipulation

Eric Caragata, Fabiano Pais, Luke Baton, Jessica Silva, Marcos Sorgine, Luciano Moreira

Fundação Oswaldo Cruz / Minas, Universidade Federal do Rio de Janeiro

Abstract

The dengue vector, Aedes aegypti, has recently been successfully transinfected with the wMel and wMelPop strains of the endosymbiotic bacteria Wolbachia pipientis. The obligate intracellular bacteria manipulate host biology, and interfere with dengue virus infection, but are predicted to become less pathogenic over time as they adapt to their new host's biology. Native infections, like wFlu in Aedes fluviatilis, offer a window into a potential future, where Wolbachia infections causes few if any manipulations, which may result in a loss of dengue interference. We generated the transcriptomes of naturally infected and Wolbachia-free A. fluviatilis mosquitoes to see the extent of transcriptional changes due to loss of infection. Whole mosquito transcriptomes were sequenced with an Illumina Platform. Sequences were first mapped to the available predicted transcriptome of the A. aegypti but, given the low percentage of correct mapping, reads were assembled into contigs for downstream analysis. The assemblies were then subjected to annotation steps and used as a reference for read mapping. Then, an R package was used to identify differentially expressed contigs between the two mosquito strains. Real-time PCR was performed to confirm differential expression of a set of 6 randomly selected genes candidates of A. fluviatilis. Only one gene did not show the same pattern of expression in comparison to the transcriptome. As expected, preliminary results indicated that wFlu strain affects the expression of fewer genes than either of the Wolbachia transinfections in A. aegypti. There was no evidence of increased immune gene expression, which is often associated with transinfections in mosquitoes. However, many of the transcriptional changes were related to metabolism and stress response, similar to wMel and wMelPop in A. aegypti. Common gene expression changes were linked to carbohydrate metabolism, membrane transport and cellular signaling, all likely to be products of Wolbachia exploiting host biology to guarantee its own survival. Our results indicate a similar pattern of transcriptional changes are associated with native Wolbachia infection in A. fluviatilis and A. aegypti, although on a lesser scale. Funding Agencies: CNPq and FAPEMIG.

A Toolbox for RNA-Seq Analysis

Juliana Costa Silva, Douglas Silva Domingues, Luiz Filipe Protasio Pereira, Mariangela Hungria da Cunha, Fabrício Martins Lopes

Federal University of Technology - Paraná, Universidade Estadual Paulista, UNESP, Embrapa Café/IAPAR, Empresa Brasileira de Pesquisa Agropecuária

Abstract

Application of next generation sequencing technology in cDNA sequencing (RNA-Seq), in transcriptomic studies has become largely accessible. RNA-Seq is suitable for transcript discovery, depicting mechanisms of gene regulation and differential gene expression analysis. When compared to microarrays, RNA-Seq brings some advantages: one of the most important advantage, is unnecessary reference sequences or any other prior knowledge, allowing confident results even in non-model organisms. A large number of methods were developed for differential gene expression analysis in RNA-Seq. However, there is not a consensus about the better approach based in the study design, and the better choice of software depending on the experimental aims. Typical approaches use Poison or negative binomial distributions to model the gene expression profiles and a variety of normalization procedures. This work proposes the development of an interactive software tool analysis of RNA-Seq data. This software tool will create a stream to differential expression analysis, guaranteeing the order and a pattern of process. Besides, it will allow the user enables the user choice among different techniques of analysis for each processing step, resulting in a useful toolbox for RNA-Seq analysis. It is adopted the Java technology for development in order to make possible its use in different operational systems. The data analysis provided by the proposed approach is composed by three main steps: i) Quality: provide information about the reads quality, and reads mapping quality against the reference genome (if existent) using Bowtie2 or BWA. Low quality mapping sequences could be removed using PicardTools software; ii) Mapping and quantification: assemble can be performed with Cufflinks or Trinity (de novo); in case of reference-based analysis the quantification BaySeq, Cuffdif, edgeR or HTSeq can be used; iii) Enrichment: It available annotation in public databases such as Uniprot. We expect to produce a software tool for RNA-Seq data analysis as well as the spread of these analysis techniques. We consider that access to this toolbox can facilitate comparative transcriptomic studies and help simplified its use in Bioinformatics classes of RNA-Seq analysis.

Transcriptome profiling in Leishmania amazonensis promastigotes associated with virulence attenuation

Gabriela Flavia Rodrigues-Luiz, Mariana Costa Duarte, Daniel Menezes-Souza, Ricardo Toshio Fujiwara, Eduardo Ferraz Coelho, Daniella Castanheira Bartholomeu

UFMG

Abstract

Leishmaniasis is one of the most important neglected tropical diseases, affecting mainly the poorest people in developing caused by obligate intracellular parasites of the genus Leishmania. It is known that in vitro cultivation of Leishmania spp. for long periods results in a progressive loss of virulence. After 10 years of publication of the first complete genome of Leishmania genus, advances in sequencing technology have provided a large accumulation of genomic, transcriptomic and proteomic data of these taxa. Thus, the focus of this work was to integrate -omics data with bioinformatics resources to contribute to a better understanding of an important biological aspect of this parasite: the loss of virulence after successive periods of in vitro cultivation. For this purpose, we evaluated the difference in expression profile of L. amazonensis promastigotes freshly isolated from experimentally infected mice (R0) and parasites that were cultured after 30 in vitro passages (R30) by RNA-seq. In vitro macrophage infection confirmed the reduction of infectivity from R0 to R30 samples. In this transcriptome study, we have identified 626 genes with significant differential expression, 66.13

StreptoRNAdb: Database system integration of ncRNAs Streptococcus strains

Tatianne C. Negri, Ivan Rodrigo Wolf, Laurival Antonio Vilas-Boas, Alexandre Rossi Paschoal

UTFPR - Universidade Tecnológica do Paraná, UEL - Universidade Estadual de Londrina

Abstract

Group B Streptococcus agalactiae (GBS) are responsible for great economic losses in milk production, aquaculture and for serious bacterial infections in humans. Consequently, they have been isolated from different sources as domestic animals, humans, horses, monkeys, cattle and fish. Although research shows that non-coding RNAs (ncRNAs) act as gene expression modulators, just a few candidates have been described for GBS. To fill this gap, here we described the StreptoRNAdb – a database system that integrates the non-coding RNAs bioinformatics meta-analysis results from our researcher group. A hybrid bioinformatics approach was used for identification and characterization of ncRNAs in nine GBS genomes from different isolation sources: human, milk and fish. This hybrid ncRNA approaches includes: INFERNAL, nocoRNAc, sRNAscanner, similarity search against ncRNA sequence public databases described in NRDR (The Non-coding RNA Databases Resource). The results showed an average of 56 candidates per genomewhich 38 are found in conserved regions of all lineages and 2 in unique regions of each genome. The analysis revealed that virulence genes are among putative targets of these candidates. The ncRNAs results from nine GBS genomes (NEM316, 2603V/R, A909, 09mas018883, ILRI005, ILRI112, SA20-06, 2-22, GD201008-001) was integrated in a unique and easy database system. The database was modeled based on the bioSQL which can be implemented in MySQL database. The website is now in development step in PHP language. By using StreptoRNAdb the user will have a user friendly website, which contains information about ncRNAs, target genes and expression data on public RNA-Seq data. The Website allows user search by simple or advanced search, by genomic location and also a graphic visualization based on circos. All these database systems are public available, allowing the scientific community explore the non-coding RNAs research performed from our group.

Homology-based annotation of non-coding RNAs in the genomes of Schistosoma species

Eddie Luidy Imada, Glória Regina Franco

Universidade Federal de Minas Gerais

Abstract

Schistosomiasis is a disease caused by species from Schistosoma genus, which affects around 210 million people worldwide. Recently, non-coding RNAs (ncRNAs) have been recognized as major players in several cellular mechanisms, such as regulation of gene transcription, chromatin and DNA metabolism, RNA stability and modification and protein synthesis. However, very little is known about ncRNAs in the Schistosoma genus. While tRNAs, rRNAs and other ncRNAs, such as snRNAs and snoRNAs can be easily recognized, the genomic prediction of other classes of ncRNAs, in special, the long ncRNAs is still a challenge. Finding similarity among ncRNAs can be a complex task, as some ncRNAs are very poorly conserved at sequence level in distantly related species. Accordingly, this challenge has led to the development of several computational approaches to functionally annotate ncRNAs, such as the automatic learning of statistical models. Here we present the development of a scalable approach by using tools to perform multiple genome alignment and statistical models to detect conservation at sequence and structure level of ncRNAs. We applied this pipeline to genomes of three species of the genus Schistosoma downloaded from SchitoDB. The initial prediction pipeline consisted in the detection of sequence conservation in intergenic and intronic regions using the MAUVE alignment tool followed by assessment of coding potential by CPAT and structural conservation detection using Infernal (based on the Rfam database). In addition, in house PERL scripts were produced to parse, adapt and integrate these analyses and facilitate the reporting and interpretation of results. A total of 249.005 intergenic and intronic sequences were found to be conserved between the analyzed genomes, whereas 4511 of these sequences presented structural homology with known RNA families in Rfam. Further steps adding transcriptomic data and downstream analyses to validate the presence of transcripts are in process. This approach constitute the first step to a better understanding of the non-coding genome of these human parasites.

Identification and characterization of microRNAs and their targets genes in the human parasite Echinococcus canadensis

Natalia Macchiaroli, Lucas Maldonado, Marcela Cucher, Laura Kamenetzky, Mara Rosenzvit

Instituto de Investigaciones en Microbiología y Parasitología Médica, IMPaM (UBA - CONICET)

Abstract

MicroRNAs (miRNAs), a class of small non-coding RNAs, are key regulators of gene expression at post-transcriptional level and play essential roles in biological processes such as development and metabolism. MiRNAs silence target mRNAs by binding to complementary sequences in the 3'untranslated regions (UTRs) of their target mRNAs. Here, we perform a comprehensive analysis of miRNAs and their targets in the cestode parasite Echinococcus canadensis, one of the causative agents of the neglected zoonotic disease cystic echinococcosis. Small cDNA libraries from two developmental stages, protoscoleces and cyst walls of E. canadensis were sequenced using Illumina technology. For miRNA prediction, miRDeep2 core algorithm was used. Differential expression analysis of miRNAs between developmental stages was estimated with DESeq and validated using poly-A RT-qPCR. Furthermore, a high confidence set of 3'UTRs were predicted. Potential mRNA targets of differentially expressed miRNAs were identified using miRanda algorithm and then filtered using different criteria such as conservation of miRNA binding sites in orthologous mRNAs of other cestode parasites in order to obtain a high confidence set of predictions. Functional information of the potential mRNA targets was obtained from GeneDB, wormParasite DataBase and KEEG databases. In this study we used a high-throughput approach to expand the miRNA repertoire of E. canadensis. Differential expression analysis showed highly regulated miRNAs between life cycle stages, suggesting a role in maintaining the features of each developmental stage or in the regulation of developmental timing. Here we confirmed the remarkable loss of conserved miRNA families in E. canadensis, reflecting their low morphological complexity and high adaptation to parasitism. We performed the first in-depth study profiling of small RNAs in the zoonotic parasite E. canadensis. We found that miRNAs are the preponderant small RNA silencing molecules, suggesting that these small RNAs could be an essential mechanism of gene regulation in this species. Functional analysis of the differentially expressed miRNAs and their potential targets will contribute to elucidate their role in the parasite biology. MiRNAs associated with parasite development, metabolism, host-parasite interaction and survival represent potential targets for the development of new therapeutic interventions.

The impact of spliced leader trans-splicing processing in the predicted proteome of Schistosoma mansoni

Jéssica Hickson, Mariana Boroni, Rennan Moreira, Michele Pereira, Willian Prado, Carolina Borges, André Reis, Mainá Bitar, Andrea Macedo, Carlos Renato Machado, Glória Franco, JS Hickson, M Boroni, RG Moreira, MA Pereira, WS Prado, CS Borges, ALM Reis, M Bitar, AM Macedo, CR Machado, GR Franco

Departamento de Bioquímica e Imunologia, UFMG, BH, MG, Brazil, Instituto Nacional de Câncer, RJ, Brazil, Departamento de Biologia Geral, UFMG, BH, MG, Brazil

Abstract

Spliced leader trans-splicing (SLTS) is a process for pre-mRNA maturation in which a small exon (spliced leader - SL) is incorporated to the 5′ portion of pre-mRNAs. It has been shown that the SLTS mechanism acts in every life-cycle stage of the parasite Schistosoma mansoni, processing a great variety of transcripts. Nevertheless, the impact of SLTS processing in the proteins of this organism has not been described yet. At the present study, RNA-Seq libraries of different stages of S. mansoni were submitted to a Bioinformatics pipeline, with the objective of describing the possible consequences of SLTS processing at the protein level. We are currently analyzing a subset of the available data, representing the adult worm library, sequenced by the Ion Torrent platform. Cutadapt was used to string trim SL-containing transcripts. Trimmed RNAs were aligned to the reference genome with the program Bowtie2. A total of 555 transcripts processed by SLTS were identified, amongst which 67 had more than one identified splicing site. From these, we have translated the nucleotide sequence of 87 transcripts using EMBOSS Transeq and further investigated possible shifts of open reading frames. We subsequently analyzed all changes in the translated sequence resulting from alternative SL incorporation using CDS (Conserved Domain Search), in order to identify functional consequences that may arise from the exclusion or inclusion of sequence elements such as uORFs, signal peptides, regulatory regions, etc. From all analyzed proteins, in comparison with their original coding sequences, 18 retained all functional domains, 22 lost all functional domains, 38 lost part of their functional domains and 9 are hypothetic proteins with no prediction available. These alternative peptides generated by SLTS processing may still be functional or even perform different functions, suggesting a biologically meaningful role for the SLTS mechanism. Further studies will help us to shed light on the dynamics of alternative SL incorporation across the life-cycle stages of S. mansoni.
Supported by: CAPES, CNPq and FAPEMIG

Evolutionary aspects of gene duplication in Drosophila

Mariana Kanbe, Nicholas VanKuren, Maria Vibranovski

Biomedical Sciences Institute, University of Sao Paulo, University of Chicago, Institute of Biosciences, University of Sao Paulo

Abstract

Duplication, a frequent molecular mechanism for generating new genes, is a mutational process that results in two copies of the same genetic information. Therefore, due to accumulation of deleterious mutations, the degeneration of one of the copies is the most commmon outcome. Rarely, however, both copies can be mantained in a population if one of them acquire a new function by accumulation of adaptive mutations, a process named neo-functionalization. Drosophila retrogenes, generated by retrotransposition duplication, occurs more often from the X-chromosome to autosomes. Many studies showed that such retrogenes exhibit male-biased expression pattern, being preferably expressed in testis in contrast to their broadly expressed parental genes. Likewise the observed X chromosome paucity of male-biased genes (demasculinization), different hypotheses based on natural selection have been proposed to explain those results such as male meiotic sex chromosome inactivation, sexually antagonistic selection and dosage compensation. Those hypothesis predicts that X chromosome is not a favorable genomic region for male expression. However, a recent study (Metta and Schlötterer, BMC Evol Biol 2010, 10:114) proposed that retrogene movement out of the X chromosome is an intrinsic property of retrotranposition rather than a product of adaptive process. Their hypothesis is based on the observation that even retrogenes where parental copy has degenerated mantained the same expression profile, not sex-biased expressed, as the original copy. However, in this study, whole body of male and female flies were used for differential sex-biased expression determination, a method known to have less power to detect male and female-biased genes than gonad comparisons. Recently, our group has generated ovary and testis transcriptome using next generation sequencing for three Drosophila species: D. melanogaster, D. ananassae and D. pseudoobscura (VanKuren and Vibranovski. J.Genomics 2014; 2:64-7). Here, we performed differential expression analyses of gonadal data to test if retrogenes moving from the X-chromosome show different expression than their parental gene. Our data show that, in general, gonadal expression detect significant more sex-biased expression which could lead accuracy, sensitivity and power of detection.

OVERACTIVE GENES: A NEW CONCEPT FOR TISSUE-SPECIFIC GENES

Lissur Azevedo Orsine, Henrique Assis Lopes Ribeiro, Glaura Conceição Franco, José Miguel Ortega

Federal University of Minas Gerais

Abstract

In general, transcriptomics studies use an expression threshold to determine if a given gene is special to a given tissue. For some genes, expression oscillates around low levels, while for others, around high levels. Thus, a gene can be expressed under the cutoff and still be relevant or evenly expressed in many tissues but highly expressed in one. Recently RNAseq data for several tissues have been made available. Observation of expression data suggests that when a gene does not have tissue-specific function, its expression varies around a mean. Besides, when the gene is important for the tissue, there is a distinct of expression in that tissue. We propose the concept of overactive genes: genes whose expression is outlier in those tissues in which they are important. In the first experiment, we extract the outliers of expression data obtained from the FANTOM5 project using a script in R. The expression of all 16,397 genes in 56 tissues was normalized to TPM (transcripts per million). In the second experiment, the expression of 18,996 genes in 32 tissues provided by Uhlen Lab was analyzed following the same experiment design. These platforms presented, respectively, 1975 and 3728 genes that were non-overactive in any tissue. A set of housekeeping genes that were not outliers and which expression is correlated (Pearson test) was used for further normalize the expression, and the non-overactive genes determined were 1675 and 2886, respectively. The validation was made through a case study, in which the placenta overactive genes were compared with placenta-related genes annotated in Gene Ontology. In this case study, a group of 525 overactive genes were identified using both platforms and both normalizers. In this group, 17 genes had annotated with placenta processes in Gene Ontology. In summary, the concept of overactive genes depicts genes that might play seminal functions specifically in some tissues. Supported by: CAPES and FAPEMIG.

Analysis of the Trypanosoma cruzi coding transcriptome in response to gamma radiation by high-throughput RNA sequencing

Michele Pereira, Priscila Grynberg, Mariana Boroni, Helaine Graziele Vieira, Dominik Kaczorowski, Andrea Macedo, Carlos Renato Machado, John Mattick, Glória Regina Franco

UFMG, Embrapa Recursos Genéticos e Biotecnologia, INCA Garvan Institute of Medical Research,

Abstract

Trypanosoma cruzi, the etiologic agent of Chagas disease, is a kinetoplastid organism highly resistant to DNA damage caused by ionizing radiation. After a dose of 500 Gy of gamma rays, the genomic DNA is fragmented. Interestingly, the parasite is able to restore the chromosomal bands pattern in less than 48 hours. Previous studies using microarrays and 2D PAGE followed by MS/MS analyzed how gamma rays affect T. cruzi gene expression. Microarray analysis showed that transcripts related to basal metabolic functions were down-regulated. In contrast, the up-regulated category was mainly composed by obsolete sequences, hypothetical proteins and Retrotransposon Hot Spot genes. Proteomic analysis indicated that active translation is essential for the parasites recovery. The presence of shorter protein isoforms after irradiation suggests the occurrence of post-translational modifications and/or processing in response to gamma radiation stress. Our study aims to analyze the gamma radiation effect on the T. cruzi coding transcriptome by RNA-seq. Epimastigote cells from CL Brener strain were exposed to a dose of 500 Gy of gamma rays. Total RNA was extracted from non-irradiated (control) and irradiated cells (4, 24 and 96 hours post-irradiation). Two biological replicates were produced for each condition. RNA-seq paired-end strand specific libraries were prepared using poly(A) enrichment/dUTP incorporation and sequenced on Illumina HiSeq 2500 platform. Reads were edited in silico to remove rRNA sequences and adapters/low quality bases, using SortMeRNA and Trimmomatic softwares, respectively. All samples and biological replicates were combined into a single RNA-Seq data set and assembled by Trinity (25-mers) to generate a single reference transcriptome assembly. A total of 113,477 transcripts (71,446 genes) were generated with an average contig length of 710.68 bp. Reads from each sample were aligned to the reference transcriptome and the estimated transcripts expression level for each sample was obtained by RSEM. Differential expression analysis was performed using DESeq2. 182, 368 and 876 transcripts were down-regulated when 4, 24 and 96 hours post-irradiation samples were compared with WT samples, respectively. On the same way, 94, 286 and 729 transcripts were up-regulated. Variance stabilizing transformation was used to cluster samples with a heatmap. Our results showed that irradiation affect gene expression in a time-dependent manner, the same observed in microarray and proteomic analyses. GO enrichment, transcriptome annotation and pathways analysis will further be performed. This study will help to understand how the parasite can handle such a harmful stress and how gene expression is coordinated to maintain its survival.

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

Tiago Bruno Rezende de Castro, Mariana Boroni, Nayara Toledo, Neuza Antunes, Afonso da Costa Viana, Carlos Renato Machado, Egler Chiari, Glória Regina Franco, Andrea Mara Macedo

Federal University of Minas Gerais, National Cancer Institute

Abstract

Even 100 years after Chagas disease discovery in 1909, the molecular bases of a variable tissue distribution of Trypanosoma cruzi strains in their mammalian hosts remains to be elucidated. Our group has previously shown that different strains of T. cruzi (JG and Col1.7G2) had a differential tissue tropism in BALB/c mice upon infection. The JG strain prevailed in the heart, while Col1.7G2 was mostly present in the rectum of mice. However, using a mixture of equivalent amounts of both strains in C57BL/6J mice, the tissue distribution was the opposite. Nevertheless, C57BLKS/J congenic mice (containing the MHC region of BALB/c mice but genetic background of C57BL/6J mice) infected with a mixture of the two strains showed the same distribution pattern of the BALB/c mice, evidencing the importance of both, parasite and host, in determining the disease course. Seeking to elucidate which host genes could be involved in this phenomenon, we have evaluated the gene expression profile, by RNA-Seq, the hearts of BALB/c mice infected with either JG, Col1.7G2 or a mixture of both strains. At 15 days after infection, representing the acute phase, mice were sacrificed and their hearts submitted to total RNA extraction. RNA was sequenced using the Illumina platform HiSeq 2000. Resulting RNA sequences were aligned to the mouse reference genome using Tophat, and counts of reads per gene were obtained using HTSeq. Downstream analyses were performed using Bioconductor packages. The Gene Ontology annotation tool DAVID was used to categorize genes in biological processes. In total, when compared to the non-infected group, 1,180 differentially expressed genes (DEGs) were detected in the group infected with Col1.7G2, 1333 in the group infected with JG and 2,260 in the group infected with the mixture of strains. From these genes, only 15% were common to all infected groups. When mapping DEGs to mouse chromosomes, we could detect that bands where MHC region is located, had the highest number of DEGs in comparison to all other chromosomal bands. Interestingly, Gene Ontology analyses shown that the majority of the downregulated genes were related to the host immune response to infection. On the other hand, upregulated genes were mainly involved in energy metabolic pathways and protein translation. Taken together, our data suggest that the differential distribution of T. cruzi among tissues could be caused by the different strategies used by each strain to survive in the infected the host. Financial support: FAPEMIG and CAPES

Microarray gene expression analysis of neutrophils from elderly septic patients

Diogo Vs Pellegrina, Patricia Severino, Marcel Cerqueira Machado, Fabiano Pinheiro da Silva Eduardo Moraes Reis

Universidade de São Paulo, Hospital Israelita Albert Einstein, Faculdade de Medicina da Universidade de São Paulo, Instituto de Química, Universidade de São Paulo

Abstract

Sepsis is one of the highest causes of mortality in hospitalized people, and its incidence is likely to increase substantially with age. Despite its increased prevalence and mortality in older people, their immune responses appear similar to that in younger patients. The purpose of this study was to conduct a genome-wide gene expression analysis of circulating neutrophils from old and young septic patients and matched controls, to better understand how aged individuals respond to infection. Expression profiles from 12 septic patients and 12 healthy controls (6 adults and 6 elders in each group) were collected using 60K element microarrays that interrogate 29402 protein-coding and 17610 noncoding RNAs. After data filtering, the samples expression were clearly distinguished by septic status or age using unsupervised clustering algorithms. Next, the differential gene expression was estimated using two different algorithms (RankProduct and SAM), and only those detected by both methods were considered significantly deregulated. Those gene could be used to differentiate immune responses of the elderly from those of young people. This data was used as input to Ingenuity's IPA software to access its enriched pathways. Our results identified major molecular pathways that are particularly affected in the elderly during sepsis, which might have a pivotal role in worsening clinical outcomes compared with young people with sepsis. These included genes related to oxidative phosphorylation, mitochondrial dysfunction and TGF-β signaling, among others. Later a q-PCR validation experiment was performed, to confirm the expression of genes from our most highlighted pathways. Work supported by FAPESP and CNPq.

Gene correlation networks with dual RNA-seq (Dual-seq) data

Caio Godinho, Michael Sammeth

Federal University of Rio de Janeiro (UFRJ)

Abstract

Sequencing the RNA population of two interacting organisms simultaneously (Dual-seq), e.g. in a pathogen-host model, is not a completely new endeavour to scientists. However, the growing power of sequencing platforms is changing the way it is done, creating new challenges to bioinformatics. It is possible now to sequence in the same protocol all sizes of RNA molecules, from mature miRNA to mRNA, with enough depth for their detection, eliminating the step of physical separation of the interacting organisms. Dealing with read normalization, considering their very different lengths (ranging from tens to hundreds nucleotides), and also their distinct sources - in this case the genomes proportions on the sample - is one of those challenges. Inferring gene co-expressions and understanding the intricate network of this relationship is another one of them. In this work we consider the mitochondria-nucleus relationship of human tissues as an example of two interacting genomic units, using this model as a training set to upcoming experiments. With this approach we address two key challenges of Dual-seq and gene co-expression networks. The first of them is the determination of the number of replicates necessary for building a reliable co-expression network between two different genomic units. Secondly, the understanding of how distinct concentrations of one genome copy interferes and modulate this molecular crosstalk. Altogether, with our results and analysis, we aim at contributing to the establishment of new and adequate protocols for Dual-seq experiments, and to a better interpretation of the complex molecular interplay underlying this coupled genome interaction.

Comparative analysis of miRNAs in Dipteran insects

Karla de Oliveira, Eric Aguiar, Flávia Ferreira, Roenick Olmo, Jean-Luc Imler, João Marques

UFMG, CNRS

Abstract

Comparative analysis of miRNAs in Dipteran insects MicroRNAs (miRNAs) are small non-coding RNAs (ncRNAs) involved in post-transcriptional regulation of gene expression. We observed high conservation of components of the miRNA pathway among dipteran insects. However, features of miRNAs themselves are as important as protein components of the pathway. The analysis of miRNAs in different organisms has greatly advanced due to high throughput sequencing technologies but it is somewhat limited by the availability and quality of reference genomes. Here, we deep sequenced small RNA libraries from sand flies (Lutomyia longipalpis), mosquitoes (Aedes aegypti) and fruit flies (Drosophila melanogaster) to evaluate overall features of miRNAs in these three insects to identify patterns of conservation or divergence. Since sand flies miRNAs were still unannotated, we first identified Lutzomyia miRNAs using miRDeep2 followed by manual curation. The same strategy was applied to Aedes samples but using already annotated miRNAs as a reference. We noted that approximately 20% of the highest expressed miRNA in fruit flies were not identified in sand flies and mosquitoes, which would seem to suggest that they are not conserved. However, considering aspects of miRNA pathway conservation and quality of assembled genomes in Aedes and Lutzomyia, we decided to use Drosophila miRNA precursors as reference to try fish out possible high conserved miRNAs. Indeed, using this strategy, we were able to identify miRNAs in Aedes (13) and Lutzomyia (23) that are conserved compared to Drosophila but their corresponding loci are absent from reference genomes. After this analysis, we observed 69 unique miRNA precursors and 109 mature miRNAs that were conserved in the all three insects, representing 80% of unique precursors in sand flies and 61% in mosquitoes. Conserved miRNAs presented similar characteristics such as origin within the precursor (5p or 3p), the size profile (21-23 nt) and enrichment of U at the 5' end. We also observed high correlation of miRNA expression between Drosophila and Aedes (0.79), Aedes and Lutzomyia (0.80) and Lutzomyia and Drosophila (0.69). In summary, our work emphasizes the overall conservation of miRNAs and how it can be explored to allow identification of miRNAs in organisms whose genome is poorly assembled or not available. Our results also show that miRNAs that have the same seed sequence also have similar expression levels and origin. This suggest that once miRNAs acquire an important function, there is pressure for conservation of not only the sequence but also its origin and expression.

Bioinformatic analysis of RNA-Seq data to search for novel prognostic/diagnostic biomarkers of pancreatic ductal adenocarcinoma

Omar Julio Sosa, Vinicius Ferreira da Paixão, João Carlos Setubal, Eduardo Reis

Universidade de São Paulo (USP)

Abstract

Pancreatic ductal adenocarcinoma (PDAC) is one of the most deadly human malignancies. The only curative treatment available is the surgical removal of the tumor in early stages of the disease. Current methods for early detection and treatment are poor, justifying more studies in this field. We aim to generate a high-resolution catalog of the PDAC transcriptome, to reveal transcriptional alterations associated with the malignant phenotype of PDAC and point to novel disease biomarkers. To that end we are implementing an informatics pipeline to detect and evaluate the differential expression in PDAC of well-annotated genes, including long noncoding RNAs (lncRNAs). In addition, we will search for novel lncRNAs and alternative splicing isoforms of protein-coding genes expressed in pancreatic tissues. In our work 28 patient matched samples of PDAC and nontumor adjacent pancreatic tissue were processed for the production of cDNAs libraries using the TruSeq Stranded Total RNA with Ribo-Zero Gold Kit, and then sequenced with Illumina HiSeq 1500 Platform (Butantan Institute, São Paulo) producing a mean of 17 million 100 nt paired-end reads per sample. The initial quality control of reads was made with FastQC, and low quality bases were trimmed (Trimmomatic). After the alignment with the reference human genome (TopHat2), a count table with the number of reads per gene for each sample was created using HTSeq. For evidence of changes across experimental conditions we use DESeq2, identifying 310 genes up-regulated (p-adjusted<0.001 and lfc>3.32) and 354 down-regulated (p-adjusted<0.001 and lfc<-3.32), comprising 156 lncRNAs. The differential expressed genes detected are involved in pathways related to cancer, regulation of cell cycle process, membrane and secreted proteins, indicating an opportunity for the identification of novel putative biomarkers for the diagnosis of the disease. In addition of these promising results, we plan to look for novel genes and isoforms using a combined strategy of de novo and with reference assembly. Work supported by FAPESP, CNPq and CAPES.

Reference-based and de novo assembly as a combined strategy to identify canonical transcripts and potential novel splice variants in proteogenomics

Raphael Tavares, Nicole de Miranda Scherer, Carlos Gil Ferreira, Fabio Passetti

Oswaldo Cruz Foundation (FIOCRUZ), Instituto Nacional de Câncer

Abstract

Transcriptome assembly from RNA-Seq data has been a challenging field, especially in humans where approximately 90% of genes produce more than one transcript due to alternative splicing events and it is expected that most of them result in alterations on the polypeptide chain. In conjunction with the emerging field of proteogenomics, the transcriptome assembly may be a promising strategy to obtain the comprehensive proteome from an organism. Reference-based and de novo assembling methods have been developed and used in the transcriptome analysis for many organisms. Here, we used a strategy combing reference-based assembly to detect canonical transcripts and de novo assembly to identify potential novel splice variants with focus on proteogenomics. mRNA-Seq reads from mouse liver publicly available data (SRA ID SRR1462347) were aligned against mouse genome (mm9) and submitted to a reference-based assembly with Cufflinks. We filtered reads to obtain only those discarded in the reference-based assembly, which were aligned to known mouse gene sequences to cluster related reads. Each cluster was then separately used by the Trinity de novo assembler for isoform detection. Mouse liver in fetal, new born and day one stages were used to test our approach, reconstructing 8,185, 8,027, 11,239 de novo transcripts from 1,122, 1,041, 1,178 genes, respectively. This preliminary results indicate that the combination of reference-based and de novo assembly are viable and a promising strategy to explore transcriptome assembly with focus on proteogenomics. Our next step will consist to use this data as input to the in silico transcriptome translation method developed by our group termed, SpliceProt.

Revealing the MAPKs signaling pathways in Schistosoma mansoni by transcriptome analyses

Sandra Gava, Naiara Paula, Fabiano Pais, Anna Christina Salim, Flávio Araújo, Guilherme Oliveira, Marina Mourão

CPqRR, Instituto Tecnológico Vale

Abstract

Eukaryotic protein kinases (ePKs) are key regulators of cellular function and act phosphorylating transcription factors causing changes in gene expression, thus in cellular behavior. The characterization of mechanisms and molecules involved in cell signaling are essential to understanding the biology of the Schistosoma mansoni and its ability to adapt to different hosts. Additionally, Schistosoma ePKs are proposed as potential targets for the development of new anti-schistosome drugs. The S. mansoni genome encode 252 ePKs, corresponding to 2% of the predicted proteome, nonetheless, only 24 have any experimental evidence. Due to rare data availability, the main motivation of this study is to contribute to ePKs experimental characterization by the identification of genes target of regulation of MAPKs and its phosphorylation counterparts. Therefore, five selected genes from the MAPK signaling pathway (SmCaMK2, SmERK-1, SmERK-2, SmJNK, Smp38) and unspecific (GFP) and negative controls were depleted by RNA interference in schistosomula, including three biological replicates. After two days of culture, total RNA extraction, cDNA synthesis, and quantitative PCR (RT-qPCR) was performed. For all genes selected, we observed approximately 75% reduction on transcript levels, excluding SmERK-2. RNA-Seq libraries were prepared with RNA derived from knockdown parasites according to the Truseq stranded mRNA Library Prep protocols and were sequenced on Illumina HiSeq 2500 platform. Were generated 11 paired-end libraries containing reads of 100 bp, ranging from 47 to 184 million reads per library, with GC content of 38-39%. The sequences were aligned to the latest S. mansoni reference genome (version 5.0) using TopHat2 (version 2.1.0). The resulting alignment files are being used as the input for the gene finder Cufflinks. Cuffdiff and Deseq2 will be used to calculate the difference in expression of transcripts in different treatment conditions. Thus, we aim to expand the knowledge around the genes target which are regulated by SmCaMK2, SmERK-1, SmERK-2, SmJNK and Smp38 revealing the MAPKs signaling pathways. This work will enable a better understanding of these pathways, elucidating functional roles of ePKs in parasite survival, reproduction and adaptation.

Metabolomics of sugarcane leaves along two experimental fields for maturation cycle study

Davi Inada, Leonardo Villela, Carolina Lembke, Milton Nishiyama-Jr, André Fujita, Glaucia Souza

Institute of Chemistry, USP, Butanta Institute, Institute of Mathematics and Statistics, USP,

Abstract

Sugarcane is a plant member of the Poaceae family and it has an important contribution for Brazilian economy. It is mostly used as a renewable source for bioethanol and sugar production. Currently, only hybrid cultivars are used in the crops for industrial production. Different molecular studies have been applied to sugarcane in Brazil. The SUCEST (Sugarcane Expressed Sequence Tags) project has built a vast EST (Expressed Sequence Tag) library that was used in different studies to better understand the underlying processes involved in gene regulation. The sucrose accumulation involves many biological processes and thus is not completely known. The understanding of these processes may aid the development of novel biotechnological procedures to increase production of crop. To improve the knowledge about this process, our research group started a metabolomic study with the SP80-3280 variety to identify all possible metabolites that are contributing to the regulation of a specific metabolic pathway. The current study is focused on analyzing the plant maturation cycle in two experimental fields, in which collected tissues from leaf +1 were used. Data were generated using a LC-ESI-TOF MS approach with two different columns (Hilic and C18) and both ionization modes (positive and negative). For the bioinformatics approach, we adopt a compound annotation pipeline using a combination of open source software packages provided by the R-Bioconductor project (IPO, XCMS, CAMERA and ProbMetab). The results found by an automated annotation were manually curated using each EIC (Extracted Ion Chromatogram) as a reference. After manual curation, a total of 119 compounds were predicted, but considering the ambiguity of the LC-MS data we obtained a total of 196 compound candidates. Hereafter statistical analysis will be applied to the data in order to identify the most significant compound and the altered pathways during the plant maturation cycle. Furthermore, a transcriptome analysis was done with leaves +1 from both experimental fields and an integrative analysis of metabolomics and transcriptomics will be performed. We intend to use these information to develop a transcriptome and metabolome data integration pipeline that will help us with the definition of major target networks and significant genes to be manipulated for improving plant productivity.

Finding new genes of lignocellulosic biomass degradation using genomics and transcriptomics analysis of the lower termite Coptotermes gestroi

Luciana Souto Mofatto, João Paulo Lourenço Franco Cairo, Melline Fontes Noronha, Ana Maria Costa Leonardo, Fabio Marcio Squina, Gonçalo Amarante Guimarães Pereira, Marcelo Falsarella Carazzolle

UNICAMP, CNPEM

Abstract

The search for renewable and sustainable energy products has increased recently, in order to substitute scarce and pollutant sources from petroleum oil and coal. Green Chemistry is the development of new chemical products that do not cause damages in the environment. One example of Green Chemistry is the bioproduct derived from microorganism's fermentation of hydrolyzed biomass, which transforms plant cell wall polysaccharides into fermentable sugars. In order to make this process economically viable, the identification of new efficient enzymes for biomass degradation is a fundamental step. Thus, termites are very interesting insects to mining these enzymes, because they live in symbiosis with bacteria, protozoa and fungus inside their guts, having the ability to degrade approximately 90% of plant-dry matter in tropical forest, converting lignocellulosic materials into sugar. In this context, the application of Omics approaches (genomic and transcriptomic) and bioinformatics analysis have important roles for searching genes for industrial and basic research purposes. Basically, the identification of enzymes can be done by integration of genomic and transcriptomic analysis using genome assembly methodology for large genome, RNA-seq data to be used in gene prediction pipeline and to identify differentially expressed genes from insect and symbiont genomes. The aim of this work was identifying genes related to lignocellulosic biomass degradation from termites Coptotermes gestroi and their symbionts submitted to different diets, such as: (1) sugarcane bagasse in natura; (2) sugarcane bagasse treated with phosphoric acid; (3) filter paper (cellulose); (4) filter paper (cellulose) and iron 3; (5) sugarcane bagasse treated with sodium chlorite and hydrochloric acid. For this purpose, we obtained genomic DNA and RNA sequencing data from these conditions. As results, Coptotermes gestroi genome was assembled using SOAPdenovo pipeline, RNA-seq reads were aligned against the assembled genome as reference for gene prediction analysis. The unmapped reads were de novo assembled using Trinity-Transdecoder-RSEM pipeline for probably finding genes from termite symbionts. All these sequences were used as reference for a second round of RNA-seq reads alignment to be used for the differential gene expression analysis. For annotation, the blast software was executed for aligning the genes against non-redundant database (NCBI), swissprot, uniref90, cdd/pfam databases. As conclusion, we found a strong relationship between termite's diet and gene expression of the symbiont organisms (bacteria, fungus and protozoa) involved on lignocellulose digestion process, mainly when the raw material is offered in natura diet. So other diets may facilitate the organism to degrade the lignocellulose biomass.

Towards a pattern recognition-based approach for sequence annotation of Phakopsorapachyrhizi

Cynara Leão Garcia, Carlos Nascimento Silla Junior, Francismar Corrêa Marcelino-Guimaraes

Universidade Tecnológica Federal do Paraná, Cornélio Procópio, Empresa Brasileira de Pesquisa Agropecuária – Embrapa

Abstract

Soybean rust caused by the obligate biotrophic fungus Phakopsorapachyrhizi is the most important biotic stress on soybean fields leading to over 80% losses. The disease has caused serious damage to soybean culture since 2001 in Brazil. Different study strategies have allowed the simultaneous monitoring of gene expression in plant-pathogen interaction, broadening our understanding of the molecular mechanisms underlying compatibility and incompatibility responses of soybean to P. pachyrhizi and thus, creating new perspectives for the development of a more durable resistance. In a previous work, the laser capture microdissection (LCM) was used to isolate the foliar mesophyll cells of rust infection sites and access site-specific processes and regulators in tolerant (compatible interaction) (BRS231) and resistant (incompatible interaction) (PI561356) soybean genotypes. RNA was extracted from the isolated cells, amplified, and sequenced with Solexaplataform. The generated paired-end sequences (54 bp) were mapped to the soybean genome and gene models for the identification of expressed genes and splicing variants. A total of 28,572 and 30,743 genes (RPKM>3) were identified for BRS231 and PI561356, respectively. The remaining reads were used to perform an ab initio assembly of P. pachyrhizi transcripts expressed at 10 dpi in planta, once the fungi genome is not available. To improve the quality of assembly, P. pachyrhizi, ESTs from Sanger sequencing reads available at NCBI were trimmed and assembled into contigs and singlets. The two assemblies were merged to form PPGC1.0 comprising 36,350 unique P. pachyrhizi sequences (unisequences) expressed at 10 dpi in planta. By combining LCM with a high performance sequencing (RNA-seq) we were able to access a potential pathogen genes expressed during the host infection and predicted a potential secretome enriched of essential genes for the infection and pathogenicity. However a large number of the pathogen contigs predicted by ab initio assembly haven't presented similarities in databases by BLAST tools. Considering the importance of the soybean rust pathogen, it is important to develop a computational tool based on pattern recognition, to carry out the mapping of the sections (no hits) that were not identified by the extracted RNAseq alignment fungus P. pachyrhizi.

Predicting non-coding RNA families based on primary sequences and secondary structures analysis using machine learning

Thaís De Almeida Ratis Ramos, Daniel Miranda de Brito, Leonardo Vidal Batista, Thaís Gaudencio do Rêgo, Vinicius Maracaja-Coutinho

Departamento de Informática, Universidade Federal da Paraíba Universidad Mayor

Abstract

Non-coding RNAs (ncRNA) are RNA molecules that do not encode proteins, but are known to provide a variety of functions in the cell. Mutations or misregulation of these ncRNAs are related to cancers, neurological diseases and neurological disorders. Like proteins, ncRNA sequences can be grouped into families; which are formed by different ncRNAs that shares a particular sequence and structural homology, resulting in a common function. Besides the variety of tools for ncRNA prediction available, those are always specific for particular classes (i.e. snoRNA, miRNA). We applied here different machine learning approaches in order to predict ncRNA families, based on characteristics from their primary sequences and secondary structures. We used Rfam database as reference, from where we randomly selected 10 sequences from 50 different families, totaling 500 sequences, with the maximum similarity between them of 80% and the size limited to 400nt. We performed tests based on the count of each nucleotide and the combination of di- and tri-nucleotides. Tests were also carried-out on resulting sequences from multiple alignments of primary sequences using Clustal Omega and secondary structures (Dot-Bracket Notation (DBN)) using MARNA. Classification tests were performed using: Naive Bayes, SMO, IBK, Multilayer Perceptron and Random Forest; through WEKA tool. In relation to nucleotides counts, the best accuracy was 57.2%. Tests using the alignments from primary sequences ranged from 82.6-89.6%; while for secondary structure alignments ranged from 86.8-92.2%. Using the combination of primary sequences and secondary structure alignments, and counting the combination of tri-nucleotides, the results varied between 85.6-92.8%. Similarly to literature, where the maximum number of tested families was 25, we performed the same tests using 10 and 20 families. The best accuracy with 10 families was using secondary structures, ranging between 98-99%; while with 20 families between 90-96.5%. Using primary sequences, secondary structures and counting the combination of tri-nucleotides together for 10 and 20 families, it was between 96-100%. Initial results obtained in this study showed that the primary sequences and secondary structures are important to the process of determining the families and function of RNAs, especially when considering that most of current available tools are exclusive for specific well-characterized classes (tRNAs, rRNAs, miRNAs, piRNAs, snoRNAs). As future work, we are extracting different attributes related to the secondary structures of ncRNAs (i.e. number of loops, number of nucleotides per loop, distance between loops, etc), and covering as many families as possible.

Expression gene levels display differences between in vivo and in vitro models

Carlos Biagi-Júnior, José Rybarczyk-Filho

Instituto de Biociências de Botucatu - Universidade Estadual Paulista

Abstract

Organisms are complexs systems composed by several of informations levels, such as: genome, proteome, metabolome, lipidome, transcriptome, etc. When one of these levels experience an external action, like physical or chemical modification or biological factor, the answer can be a reversible or irreversible efect in the system. The toxicogenomic is an area that study these effects in living organisms, using a system model as reference. Mus musculus and Rattus norvegicus are preferential model organisms before the researchers conduct analysis in Homo sapiens. The choice is justified by the similarities of these organisms to humans. An approach to conduct a global analysis of chemicals in an organism is the use of microarrays, a high sensible technique with low cost compared to RNA-seq. The Toxicogenomics allow the analysis in details of the changes in gene expression levels caused by an external stimulus in a specific organ of a speficic model organism. The mainly question for Toxicogenomics studies is if the invivo analysis can be replaced by in vitro analysis. To answer the question, we used the Japanese Toxicogenomics Project that provides microarray data from liver of R. novergicus (in vivo and in vitro) and H. sapiens (in vitro) treated with 131 drugs (approved by FDA),in differents dosages and treatment time, in a total of 20000 microarrays. In this work 5 drugs were used: aspirin, caffeine, ethanol, omeprazole and tamoxifen. The microarrays were analysed using limma package from Rstudio. We selected the differential expressed genes (log foldchange > 1 and < -1, p-value < 0.001). We construct Venn's diagram to compare several drug dosages in relation to H. sapiens in vitro, R. novergicus in vitro and in vivo. The results indicated that R. novergicus invitro showed more differential expressed genes (DEG) than H. sapiens in vitro to aspirin, caffeine and ethanol. Moreover, R. novergicus in vivo showed less DEGs than R. novergicus in vitro and H. sapiens in vitro treated with omeprazole showed more DEGS than other models. In the case of Tamoxifen, H. sapiens has only 1 DEGs, R. novergicus in vivo and in vitro have aproximatelly 11 DEGS. These results were used to built protein-protein networks to visualize the interaction among DEGs. This preliminary results allow to conclude the existence of differences among the models. The next step is use the transcriptogram methodology that connect network topologies of networks to microarray data to conduct a comprehensive evaluation of the metabolic performance.

The use of transcriptomic next-generation sequencing data to assemble mitochondrial genomes

Daniel de Andrade Moreira, Paula Cristina Cordeiro Andrade, Maithê Gaspar Pontes Magalhães, Carolina Furtado, Thiago Estevam Parente

FIOCRUZ, FIOCRUZ, FIOCRUZ, INCA, FIOCRUZ

Abstract

Current protocols to sequence mitochondrial genomes rely almost exclusively on long range PCR or on the direct sequencing. While long range PCR includes unnecessary biases, the purification of mtDNA for direct sequencing is not straightforward. We used total RNA extracted from liver and Illumina HiSeq technology to sequence mitochondrial transcripts from 34 fish species and assemble their mitogenomes. Each transcriptome was subjected to BLASTN against the complete mitogenome of the closest related species, whose mitogenome is publically available. The transcripts aligned with the reference mitogenome were used for mitogenomes assembly. Selected transcripts were edited according to the information of strand orientation given by the BLASTN result, and aligned to the reference mitogenome. A CONTIG sequence was generated for each individual fish. The CONTIG sequence was then manually checked for inconsistencies and gaps, which were filled with Ns. The mitogenomes were annotated using web-based services. Bowtie v. 1.0.0 was used to align the reads of each fish on its own assembled mitogenome, in order to estimate sequencing depth. The mitochondrial reads represented a median of 1.67% of the total transcriptome reads. Heteroplasmic sites were detected using IGV, setting the software to show positions in which the frequency of the second most frequent base was equal to or higher than 10% and the total reads number were higher than 100. We estimate to have sequenced from 91 to 100% of the mitogenomes, varying from 15,061 to 16,630 nucleotides in length. The minimum sequencing depth was 2,076x. The complete sequence of the 13 protein-coding genes and the two ribosomal RNAs were obtained for all species. Most of the 22 tRNAs were also sequenced in all species. The D-loop region was most frequently not sequenced. When sequenced, the assemble of D-loop was not trivial due to its repetitive sequence. Moreover, the use of transcriptomic data allowed the observation of the punctuation pattern of mtRNA maturation, to analyze the transcriptional profile, and to detect heteroplasmic sites. The assembly of mtDNA from transcriptomic data is complementary to other approaches and overcomes some limitations of traditional strategies for sequencing mitogenomes. This approach is faster than traditional methods and allows a clear identification of genes, in particular for tRNAs and rRNAs. This work is supported by a PEER grant from USAID (PGA-2000003446) and partially published at Gene 573 (2015) 171 – 175.

Study of the relationship between microRNAs in sex chromosomes and differential expression in autosomes of the human brain in different periods of the development.

Fátima B. Annetta, Ana Carolina Tahira, Helena P. Brentani, Ariane Machado Lima

University of São Paulo, University of São Paulo, Medical School (FMUSP)

Abstract

The elucidation of many psychiatric disorders, such as those of neurodevelopmental disorders as the Autistic Spectrum, are a current challenge. Multiple genes are involved, as well as many other genetic and non-genetic aspects, making difficult to establish a hereditary pattern. However, it is important to note that sex bias occur in many human diseases, including psychiatric disorders, and there is evidence that the sex chromosomes (X and Y) have influence on the regulation of expression on autosomal genes (non-sex chromosomes). Various associated factors have been discovered and studied, such as the role of microRNAs (miRNAs) in neurodevelopmental disorders. These small molecules have revealed an important role in the regulation of gene translation and transcription. Therefore, the objective of our project is to identify and define microRNAs present in these chromosomes (X and Y), and explore their networks associated with brain gene expression in different periods of the development. A total of 121 microRNAs were identified from sex chromosomes, using miRBase database. TargetScan was the computational tool chosen to predict the gene targets of those microRNAs. Thus, were predicted 30887 human transcripts with predicted targets sites for these miRNAs of interest. Each of these transcripts received different prediction scores based on conservation and sites types, localization and accessibility. We are working to choose a score threshold which is a reasonable balance between sensitivity and specificity in order to define which targets will be considered to further exploration. Also the integration of biological information from different databases will be performed to guide gene selection.

Large-scale analysis of transcripts processed by Spliced Leader trans-splicing in sporocysts of Schistosoma mansoni

Núbia Fernandes, Mariana Boroni, Sandra Gava, Glória Franco, Marina Mourão

CPqRR - Centro de Pesquisas René Rachou, INCA - Instituto Nacional de Câncer José Alencar Gomes da Silva, UFMG - Universidade Federal de Minas Gerais

Abstract

Schistosomiasis is a major parasitic disease, which is endemic in 76 countries. In Brazil, this is one of the most serious public health problem persisting due to the precarious living conditions in which the population is inserted. Schistosoma mansoni is the only specie described in Brazil responsible for causing schistosomiasis. This parasite is a digenetic metazoan with several unique features in its morphology, physiology and life cycle. Therefore, it is plausible a complex regulation of gene expression in S. mansoni, allowing morphological and biochemical changes that attend their physiological needs and to adapt to different environments. Therefore, the mechanism of post-transcriptional regulation, Spliced leader (SL) trans-splicing, existent in the parasite, may be important to enable these adaptations. This mechanism is only partial understood and thus a wide field for research towards the development of tools for schistosomiasis control. The SL trans-splicing occurs by the addition of a sequence identified as Spliced leader which is donated from the 5' end of a small RNA to receptor pre-mRNAs, forming the exon 5' end of mature mRNAs. Here, we carried out the identification of transcripts processed by SL trans-splicing in sporocyst of S. mansoni by constructing cDNA libraries enriched. Thus, by means of a pioneer study of transcriptomics involving the sporocyst stage, fragment libraries were constructed and used next-generation sequencing to identify 1.191 transcripts processed by SL trans-splicing in this stage. In this work, we found that 10% of transcripts expressed in the sporocyst stage are processed by SL trans-splicing when compared to the 5th version of the predicted proteome of S. mansoni and 15%, when compared to the 6,677 expressed genes identified in the transcriptome of the sporocyst stage. After the classification of transcripts in functional categories and identification of metabolic pathways we observed that the SL trans-splicing mechanism does not seem to be particularly enriched, being characterized as an ubiquitous mechanism. Together, our data improve knowledge acquired on transcriptomics studies of the parasite S. mansoni. Understanding the real function of this mechanism can assist in the future development of a therapeutic intervention tool for schistosomiasis control.

Transcriptome analysis of a murine model of melanoma progression

Flávia E. Rius, Omar J. Sosa, Vinicius F. da Paixão, Eduardo M. Reis, Miriam G. Jasiulionis

Unifesp, USP, USP, USP, Unifesp

Abstract

Melanoma is the most aggressive type of skin cancer and one of the less responsive to currently available therapy. The molecular mechanisms by which melanocytes develop into aberrantly proliferating cells, and then to a malignant tumor, are poorly known. Aiming to understand this steps that intermediate the transition of normal cells to malignant cells, we decided to investigate the transcription pattern between these stages through RNA-sequencing. For that, we used a model developed in our laboratory that consists in submitting murine melanocytes, melan-a lineage, to sustained stress conditions. These conditions consist in subjecting non-tumorigenic cells to successive cycles of anchorage blockade. It generates not only tumorigenic lineages, metastatic or not, but also lineages corresponding to intermediate steps of the process of malignant transformation. We have selected four lineages to this study: melan-a, non-tumorigenic melanocytes, 4C, pre-malignant non-tumorigenic melanocytes, 4C11-, tumorigenic non-metastatic, and 4C11+, metastatic melanoma cells. Total RNA were extracted from three different samples of these four cell lines, to produce the libraries using the TruSeq Stranded Total RNA LT- (with Ribo-Zero TM Gold) Kit. They were sequenced with Illumina HiSeq 2500 Platform (Butantan Institute, São Paulo) and generated a mean of 13 million 100 nt paired-end reads per sample. FastQC was used to verify the initial quality control of reads, and Trimmomatic to trim low quality bases. After the alignment with the reference mouse genome (TopHat2), the number of reads per gene for each sample was counted using HTSeq. To identify the differentially expressed genes between the transition lineages, we used DESeq2, considering upregulated/downregulated genes those displaying a 2-fold or greater change in expression and an adjusted pvalue <0.05. The greater difference was obtained between the non-metastatic (4C11-) and metastatic (4C11+) melanoma cell lines, with 1697 genes upregulated and 1020 downregulated. Interesting genes related to embryonic development, cell morphogenesis, cell adhesion, apoptosis, proliferation, and immune response were found in different comparisons between all the cell lines, elucidating possible mechanisms involved in melanoma progression. Since the transformation of the cells in this model is causally related with epigenetic alterations, we intend to investigate the expression of genes encoding components of the epigenetic machinery as well as regulatory long noncoding RNAs across all cell lines to better understand how both processes are connected.

MicroRNA expression during Schistosoma mansoni development

Victor Fernandes de Oliveira, Fabiano Carlos Pinto de Abreu, Roberta Versiano Pereira, Marcela Pereira Costa, Matheus de Souza Gomes, Liana K. Jannotti-Passos, William Castro Borges, Renata Guerra-Sá

Núcleo de Pesquisas em Ciências Biológicas, Universidade Federal de Ouro Preto

Abstract

Schistosomiasis is a debilitating disease that is caused by Platyhelminths of the genus Schistosoma. Due to its complex life cycle, evolutional position and sexual dimorphism, schistosomes can serve as an interesting model to investigate mechanisms of gene regulation. MicroRNAs (miRNAs) are short endogenous RNA molecules that regulate gene expression at the post-transcriptional level by targeting mRNA transcripts. Current knowledge of Schistosoma mansoni miRNAs is limited and is based on computational predictions and next-generation sequencing through adult worms libraries. In this study, we validated the expression profiles of seven mature miRNAs in different life cycle stages of the parasite using qRT-PCR and identified the miRNA target genes using a computational approach. Our results showed differential expression patterns of the miRNAs sma-miR-281; sma-miR-new_2-5p; sma-miR-new_4-5p; sma-miR-new_5-5p; sma-miR-new_12-5p; sma-miR-new_13-3p and sma-miR-new_13-5p. The computational analysis revealed miRNA target genes that are related to important biological processes, such as TGF-β signalling, glucose and lipid metabolism, tetraspanin protein (TSP), cathepsin B 1 protein (CB1) and Venom allergen-like 6 protein (VAL-6). Furthermore, most of the target genes that were found are linked to oxidative phosphorylation, suggesting that at least in part the expression of NADH dehydrogenase, cytochrome c oxidase e ATP synthase genes are regulated by microRNAs. We also observed target genes that are involved in the proteasome-ubiquitin protein degradation pathway, suggesting that miRNAs can regulate this important biological process in the parasite. Together, our results lead us to suggest that those miRNAs might play important roles in the post-transcriptional regulation of genes that are related to energetic metabolism inversion during parasite development.

Classification of Coding and Non-Coding RNAs through Random Forests: a Preliminary Analysis

Clebiano Costa-Sá, Marcelo Lauretto, Ariane Machado-Lima

University of São Paulo

Abstract

RNAs can be classified in two broad classes: coding and non-coding. Non-coding RNAs (ncRNAs) are involved in various cellular activities and also associated with various diseases such as heart attack, cancer and psychiatric disorders. The discovery of new ncRNAs and their molecular roles favours advances in knowledge of molecular biology and can also help the development of new disease therapies. Identification of ncRNAs is an active research area, and one of current approaches is the classification of transcribed sequences using pattern recognition systems based on their features. One of the classifiers for this purpose, CPC (Coding Potential Calculator), is based on SVM (Support Vector Machine) algorithm. Another robust algorithm with potential to distinguish RNAs is the Random Forests. The aim of this work is to evaluate the performance of the Random Forests algorithm for classification of RNAs in coding or non-coding, using the features specified by the CPC and also a set of new features. In order to compare the results to CPC performance, the CPC training set sequences were used, containing 5610 coding and 2670 non-coding RNAs. New features were extracted from an extractor system written in Perl. Features specified by CPC were recovered from adaptations in the CPC software. Three classifiers aggregates were created with Random Forests algorithm: the first trained with new features, the second trained with CPC features and the third trained with both sets of features. The most sensitive parameter, the number of features tested per node (MTRY), was selected from a grid, by maximizing the evaluation measures of the classifiers obtained via 10-fold cross-validation experiments. The average accuracies of these classifiers were, respectively, 96.13%, 95.52% and 96.64

Clustering algorithms application for analyzing gene expression profiles with microarray data from patients with Osteogenesis Imperfecta

Diogo Pereira Silva de Novais, Paulo Eduardo Ambrósio, Kaneto Carla Martins

Universidade Estadual de Santa Cruz

Abstract

Based on the hypothesis that genes which present similar expression profiles when exposed for some condition can be involved in related functional process or are regulated for similar cellular mechanisms, the cluster analysis between gene expression profiles can reveal important information about genes involved in some process or biological condition. However, the application of clustering algorithms has a set of uncertainties inherent to this process, once different techniques of preprocessing, different algorithms or even different parameters can reveal distinct information about the expressed genes. Thus, this work aims to analyse gene expression data from patients with Osteogenesis Imperfecta through different algorithms and to present a discussion about the identified profiles for each algorithm and possible inferences around genes involved in biological processes related to the pathology. There were analysed approximately 40000 genes expressed in cells from patients with Osteogenesis Imperfecta Type 1, Osteogenesis Imperfecta Type 3, and a healthy control set, during the process of osteogenic differentiation from mesenchymal stem cells. The results of gene expression furnished by the microarray experiment were preprocessed and the 100 genes which have the biggest standard deviation between the samples were selected to the cluster analysis. The clusters had been built through three non-hierarchical algorithms: K-means, Self-Organizing Maps and Affinity Propagation, all of them using the number of clusters as 5 and the euclidean distance as dissimilarity measure between the samples. All the algorithms have found a cluster of genes with low expression in the control samples and high expression in the subjects with the pathology, what can suggest an deviation of those genes for other process, affecting the bone tissue production. Another group of genes with high expression in the control samples and low expression in the subjects with the pathology were also found, what can evidentiate an underproduction of some gene product important for the osteogenesis. Further, the Affinity Propagation revealed a cluster of genes with high expression only in the patients with Osteogenesis Imperfecta Type 3, that possibly can show specific profiles of this pathology.

DNSAs - The de novo sequence annotation system

Marcelo Brandão

UNICAMP

Abstract

The post-genomics and next generation sequence era has bring up new challenges to the bioinformaticians. This embraces the identification and description of the non-model organisms' transcriptomic data, in special the gene ontology assignment to the de novo assembled contigs. There are other systems "on the market" but some commercial and some that can be fully used only with data from a specific assembler. Here I propose a new pipeline approach to functionally annotate these sequences independently of the assembler and up to 100 times faster than the available ones. This fastness is achieved using DIAMOND as sequence similarity proposer and the capability to distribute the analyses on a cluster environment.

Probabilistic Framework for RNA Sequence Analysis

Rafael Mathias, Alan Durham

Universidade de São Paulo

Abstract

RNA is a four nucleotides polymer denoted by A, C, G, U which represent, respectively, Anenine, Cytosine, Guanine and Uracil. The bases A and U form hydrogen bonds, as well as the bases C and G, and these kinds of base paring are called canonical. Nevertheless, other kinds of base paring can be formed. RNAs are molecules of a single string that can fold into themselves by base pairing interactions. The structure resulted from those interactions is called RNA's secondary structure. Recent studies have shown that non-coding RNAs act upon a variety of biological processes such as gene silencing, gene expression, transcription and translation control. They are also associated with various types of diseases such as cancer, neurological diseases - as alzheimer and parkinson -, cardiovascular diseases, among others. It is therefore of fundamental importance to find new non-coding RNAs and its respective secondary structure due to the close relationship between the secondary structure and the biological function of these molecules. In this work we developed a probabilistic framework using context sensitive hidden Makov models to characterize sequences and profile of sequences with arbitrary distance between symbols, such as those found in RNA sequences and RNA alignments. Our development was made as an extension of the probabilistic framework ToPS and includes optimized versions of the inference algorithms in order to achieve efficient runtimes. We compared our approach with other framworks with similar purposes and noticed that our framework proves itself quite competitive, in addition to offering increased freedom in model definition.

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

Nayara Evelin Toledo, Tiago Bruno Castro, Carlos Renato Machado, Neuza Antunes Rodrigues, Afonso Viana, Egler Chiari, Andrea Mara Macedo, Glória Regina Franco

Universidade Federal de Minas Gerais

Abstract

The parasite Trypanosoma cruzi is the causative agent of Chagas disease. Even after 100 years of its description, the elucidation of the mechanisms underlying the tissue tropism of T. cruzi in its mammalian host still is a challenge. Our group has previously shown that different population of T. cruzi (JG and Col1.7G2) had a differential tissue tropism in BALB/c mice upon infection. The JG strain prevailed in the heart, while Col1.7G2 was mostly present in the rectum of mice. However, using a mixture of equivalent amounts of both strains in C57BL/6J mice, the tissue distribution was different and Col1.7G2 was found in many tissues, including the heart. Studies of differential expression using Next Generation Sequencing data from the BALB/c infected hearts in comparison to non-infected mice showed that mice infected with Col1.7G2 had mostly downregulated genes, while animals infected with JG presented a great number of down and upregulated genes. The same phenomenon was observed for the mixture-infected group. These results demonstrate that different populations of T. cruzi can induce a distinct change in the host gene expression. Cis-Splicing is a RNA processing in which introns are removed from pre-mRNA and exons are joined to generate a mature transcript. In alternative splicing different exons and introns of the same pre-mRNA may be skipped or retained to produce different mature mRNAs, largely expanding the transcriptome repertoire. Thus, the aim of this study was to evaluate the occurrence of alternative splicing in the transcriptome of mice infected with the T. cruzi strains JG and Col1.7G2. RNASeq data were acquired from infected mice hearts using the Illumina platform HiSeq 2000. For initial analysis of the transcriptomes, the quality of sequences was accessed with the FastQC software. Subsequently, we performed alignment against the mouse reference genome using the splice-aware aligner, STAR. After alignment, the program Multivariate Analysis of Transcripts Splicing (MATS) was used for recognition of the main types of alternative splicing and for annotation of novel splice events. Our present result showed that intron retention and exon skipping prevailed in JG and mixture infected mice when compared to the control group, whereas this was not observed in hearts of animals infected with Col1.7G. In conclusion, we have shown that, in the experimental model of Chagas disease, different T. cruzi populations can significantly remodel the splicing pattern of the host and this may be relevant for disease development and differential tissue tropism.

Long non-coding RNAs in carnivorous plants: predicting lncRNAs in family Lentibulariaceae

Saura R. da Silva, Vitor F. O. de Miranda, Todd P. Michael, Daniel G. Pinheiro

Instituto de Biociências, UNESP - Univ Estadual Paulista, Câmpus Botucatu, Faculdade de Ciências Agrárias e Veterinárias, UNESP - Univ Estadual Paulista, Câmpus Jaboticabal, Ibis Bioscience

Abstract

The family Lentibulariaceae comprises three carnivorous genera: Pinguicula, Utricularia and Genlisea. The species inhabit environments that are usually deficient in nutrients, which are supplied by acquiring nitrogen and phosphorus from their prey captured by foliar traps. The family holds species with different genome sizes among the lineages (1C ranging from 61 to 1,722Mb) and even in the same species (from 60 to 131Mb, for different populations of G.aurea), and Utricularia and Genlisea genera are known to share highly increased rates of nucleotide substitution. Therefore, these plants are good candidates for future research on the complexities of plant physiology, plant nutrient pathways, evolution of angiosperm genomes and transcriptomes, which includes the mining of long noncoding RNAs (lncRNAs) and discovery of the processes related to these molecules. These mRNA-like transcripts are greater than 200nt in length, mainly transcribed by RNA polymerase II, polyadenylated, spliced, and mostly localized in the nucleus. There are increasing evidence that lncRNAs acts as important regulatory factors such as during plant reproduction, stress, developmental regulations, but few reports have examined lncRNA in non-model plants, such as carnivorous species. Considering this, the aim was to identify and characterize sets of lncRNAs in Pinguicula vulgaris, Utricularia gibba, U. intermedia, and U. vulgaris, using RNA-seq data available in GenBank. Since U.gibba and U.vulgaris transcriptomes were already available, we de novo assembled the transcriptomes of P.vulgaris and U.intermedia using high quality and trimmed reads, resulting in 172,577 and 130,537 transcript isoforms, respectively. We analyzed these sets of transcribed sequences to identify putative lncRNAs using a lncRNA bioinformatic pipeline. The lncRNA pipeline evaluates the length of transcript isoforms (>200 bp) and predicted ORFs (<100 amino acids), as well as protein-coding potential, which was based on the quality, completeness and sequence similarity to known proteins in a comprehensive database. We found 38,887, 13,983, 1,208, 20,063 predicted lncRNAs for P. vulgaris, U. gibba, U. intermedia and U. vulgaris, respectively. The pipeline discarded 12,567 to 102,570 sequences that represent ORFs, and 85 to 340 precursors to smallRNAs. These lncRNAs predictions were clustered based on sequences similarity resulting in 12 clusters with putative transcripts for at least two species, seven of them for three species and none were found for all species. Future studies will include prediction of lncRNAs in additional carnivorous genomes, assessment of whether these lncRNAs are expressed through RT-PCR, and whether they have common ancestry based on phylogenetic analyses.

Evaluation of de novo RNA-Seq assemblers in differential expression experiments

Lucas Miguel Carvalho, Zanoni Dia, Felipe Rodrigues da Silva

Unicamp, Embrapa Informática Agropecuária

Abstract

Evaluation of de novo RNA-Seq assemblers in differential expression experiments. RNA-Seq is a technology developed from Next-Generation Sequencing data (NGS) for transcriptome studies. It usually generates millions of short fragments of mRNA for contrasting treatments. Its reliability is undisputed for model organisms for which there are well assembled reference genome sequence and annotation available. However, generating an eukaryotic reference genome still is a difficult and expansive task. Assembling RNA-seq reads to describe an organism transcriptome without aligning the reads to its reference genome is called de novo transcriptomics. The objective of this study is to evaluate methodologies applied on de novo transcriptomics studies, proposing criteria for ranking the data, in order to maximize the chance of correctly identifying a differentially expressed transcript. The classification can help eliminate false positives transcripts usually found on the expensive and laborious downstream analysis, such as Real Time PCR. In this work, several parameters were tested with 3 assemblers: Trinity, Oases and IDBA-Tran. Actual RNA-Seq data from Arabidopsis thaliana and Canis vulgaris experiments were used for as input for the 3 assemblers. The list of differentially expressed genes was ranked using 15 different criteria and compared to the genes identified by the Tuxedo suite (BowTie, TopHat, Cufflinks, CummeRbund). Contrary to our expectation, the results show that the amount of true differentially expressed identified transcripts do not change significantly with reduction of input data. The assembler that consistently delivered best results was Trinity. The best ranking criteria was the transcript number of reads combined with p-value.

Predicting Piwi-interacting RNAs by deep learning

Paulo Roberto Branco Lins, Marcilio Souto, Leonardo Vidal Batista, Thaís Gaudencio do Rêgo, Vinicius Maracaja-Coutinho

Departamento de Informática, Centro de Informática, Universidade Federal da Paraíba, João Pessoa, Brazil. Centro de Genómica y Bioinformática, Facultad de Ciencias, Universidad Mayor, Santiago, Chile. Instituto Vandique, João Pessoa, Brazil. Beagle Bioinformatics, Santiago, Chile.

Abstract

Non-coding RNAs (ncRNAs) are functional RNA molecules not translated into proteins. One of the most representative ncRNA family are the Piwi-interacting RNAs (piRNAs), which ranges between 26-31 in length and interacts specifically with PIWI proteins expressed more abundantly in neuronal and germ line cells, especially in spermatogenesis. They are known to have a key role in both epigenetic and post-transcriptional gene silencing of retrotransposons. Even as the most representative ncRNA class in public databases, its prediction in transcriptome datasets is still a challenge, due to the fact that they do not possess known secondary structures or primary sequence motifs. Currently, there over 77 million piwiRNA sequences available on the public database piRBase. Due to this huge amount of data available, we aimed here to classify piwiRNAs using an architecture in which its learning takes place in a profound way on multi-layers, also called Deep Learning. For that, we used a deep learning approach with the k-mer scheme in a neuronal network with four layers (entrance layer, two hidden layers with two thousand units each, and the output layer), without pre-training the data and using Adam heuristics, on a dataset of 224,667 piwiRNAs from mouse available on piRBase. As negative control, we randomly selected the same number of mouse protein coding sequences from GENCODE. Those sequences were randomly cutted in order to have similar sizes to that of the group of real piRNAs available. Our approach predicted correctly a total of 203,385 (90.52%) real piRNAs. These results shows that deep learning approaches are useful for the prediction of new piwiRNAs. Future investigation implicates on the implementation of the pre-training and the usage of additional hidden layers, expanding the network size and exploring different architectures.

ProClaT, a new bioinformatics tool for in silico protein reclassification: case study of DraB, a protein coded from the draTGB operon in Azospirillum brasilense

Elisa Terumi Rubel, Roberto Tadeu Raittz, Nilson Antonio da Rocha Coimbra, Michelly Alves Coutinho Gehlen, Fabio de Oliveira Pedrosa

Federal University of Paraná - Curitiba

Abstract

Azopirillum brasilense is a plant-growth promoting nitrogen-fixing bacteria that is used as bio-fertilizer in agriculture. Since nitrogen fixation has a high-energy demand, the reduction of N2 to NH4+ by nitrogenase occurs only under limiting conditions of NH4+ and O2. Moreover, the synthesis and activity of nitrogenase is highly regulated to prevent energy waste. In A. brasilense nitrogenase activity is regulated by the products of draG and draT. The product of the draB gene, located downstream in the draTGB operon, may be involved in the regulation of nitrogenase activity by an, as yet, unknown mechanism. A deep in silico analysis of the product of draB was undertaken aiming at suggesting its possible function and involvement with DraT and DraG in the regulation of nitrogenase activity in A. brasilense. In this work, we present a new artificial intelligence strategy for protein classification, named ProClaT. The features used by the pattern recognition model were derived from the primary structure of the DraB homologous proteins, calculated by a ProClaT internal algorithm. ProClaT was applied to this case study and the results revealed that the A. brasilense draB gene codes for a protein highly similar to the nitrogenase associated NifO protein of Azotobacter vinelandii. This tool allowed the reclassification of DraB/NifO homologous proteins, hypothetical, conserved hypothetical and those annotated as putative arsenate reductase, ArsC, as NifO-like. An analysis of co-occurrence of draB, draT, draG and of other nif genes was performed, suggesting the involvement of draB (nifO) in nitrogen fixation, however, without the definition of a specific function.

BARHL1 is downregulated in Alzheimer's disease and may regulate cognitive functions through ESR1 and multiple pathways

Debmalya Barh, María E. García-Solano, Neha Jain, Antaripa Bhattacharya, José García-Solano, Daniel Torres-Moreno, Sandeep tiwari, Belén Ferri, Krishna Kant Gupta, Artur Silva, Vasco Azevedo, Preetam Ghosh, Pablo Conesa-Zamora, Kenneth Blum, George Perry

Centre for Genomics and Applied Gene Technology, Santa Lucía General University Hospital (HGUSL),Catholic University of Murcia (UCAM), Universidade Federal de Minas Gerais, Virgen Arrixaca University Hospital (HUVA), Universidade Federal do Pará, Virginia Commonwealth University, University of Florida, Case Western Reserve University

Abstract

AD (Alzheimer's disease) is one of the "most common forms of neurodegenerative dementia". The global prevalence of AD is as high as 24 million and in the USA alone it is approximately 5.4 million including approximately 200,000 that comprise the younger-onset AD population. According to the Alzheimer's Association, in every 68 seconds someone develops AD in the USA, which is projected to be one new case of AD in every 33 seconds by 2050. Development and maintenance of the nervous system requires coordinated actions of multiple transcription factors that may be affected in neurodegenerative disorders like AD. Reports suggest that the homeodomain transcription factor BARHL1 plays an essential role in the migration and survival of cerebellar granule cells and precerebellar neurons and its expression is upregulated during cerebellar development in human. Despite the important role of BARHL1 in brain development, no studies so far have assessed BARHL1 expression in neurodegenerative disorders like AD or Parkinson disease (PD) or in neoplastic diseases other than brain tumors. Transcription factor BARHL1 is overexpressed in medulloblastoma and plays a role in neurogenesis. However, much about the BARHL1 regulatory networks and functions in neurodegenerative and neoplastic disorders are not yet known. In this paper, using a tissue microarray (TMA), we report for the first time that BARHL1 is down regulated in hormone negative breast cancers and Alzheimer's disease (AD). Further, using an integrative bioinformatics approach and mining knockout mouse data, we show that (i) BARHL1 and ESR1 may constitute a network that regulate NTF3 and BDNF mediated neurogenesis and neural survival, (ii) this is probably linked to AD pathways affecting aberrant post-translational modifications including sumoylation and ubiquitination, (iii) BARHL1-ESR1 network may regulate beta-amyloid metabolism and memory and (iv) hsa-mir-18a, having common key targets from the BARHL1-ESR1 network and AD pathway, may regulate neuron death, reduced beta-amyloid processing, and might also be involved in hearing and cognitive decline associated to AD. We have also hypothesized why estrogen replacement therapy improves AD condition. In addition, we have provided a probable new mechanism to explain the abnormal function of Mossy fibers and cerebellar granule cells related to memory and

Mapping SOS system in Leptospira spp

Lívia de Moraes Bomediano, Renata Maria Augusto da Costa, Ana Carolina Quirino Simoes

Universidade Federal do ABC

Abstract

Leptospirosis is a tropical disease caused by pathogenic bacteria of the Leptospira genus. The disease is responsible for serious public health problems resulting in costs to the economy. The SOS response is a bacteria defense mechanism against DNA damage caused by ultraviolet radiation, high concentrations of O2 and iron inside the cell and antibiotics. This system is well studied in Escherichia coli where more than 40 genes, including recA, dinP, uvrA and recN genes were found to be involved in this kind of response and regulated by lexA and recA regulators. In Leptospira spp this response is not well characterized. This study compared in silico four leptospira genomes of interest that are already sequenced to understand the differences between pathogenic and saprophytic species of leptospira regarding the SOS system. The Escherichia coli' SOS response system sequences of 13 well studied genes in this organism were carefully annotated and used as driver sequences to the mapping of SOS system in four leptospira genomes using Blastn and Blastx searches. The comparative analysis was held using the ACT Artemis tool. The complete genome comparison between pathogenic species revealed a lower identity between their genomes than the saprophytic ones. It also showed altered genome structure, revealing a large reverse direction reading genomic reagion. The comparison between pathogenic and saprophytic species revealed even greater differences. Regarding the SOS system, the results showed that although Leptospira spp have eight SOS genes in common, the position pattern and reading direction differ between the saprophytic species and pathogenic species.

Streptococcus pyogenes serotype M1 outbreak in Brazil reveals genomic variations among lethal invasive strains

Gabriel R. Fernandes, Aulus E. A. D. Barbosa, Renan N. Almeida, Fabiola F. S. Castro, Marina C. P. Ponte, Celio Faria-Junior, Fernanda M. P. Müller, Antônio A. B. Viana, Dario Grattapaglia, Octavio L. Franco, Sergio A. Alencar, Simoni C. Dias

Centro de Pesquisas Rene Rachou - Fiocruz

Abstract

Streptococcus pyogenes, also known as group A Streptococcus (GAS), is a human pathogen that causes a large diversity of human diseases including streptococcal toxic shock syndrome (STSS). A GAS outbreak occurred in Brasilia, Brazil, during the second half of the year 2011, leading to 26 deaths. Whole genome sequencing was performed using Illumina platform. The sequences were assembled and genes were predicted for comparative analysis with emm type 1 strains: MGAS5005 and M1 GAS. The gene content was around 1850 coding genes for the invasive strains. Comparing the 4 invasive strains with references – M1GAS and MGAS5005 – we observed 1855 orthologous groups. 1505 orthologous genes were shared among all 6 genomes. Analyzing the non-core genes, we could identify 44 genes present exclusively in MGAS5005 genome. Among them we can identify some genes related to genetic processing: transposase, integrase, relaxase, transcriptional regulartor; ribosomal assembly; and energetic methabolism. In isolate Sp2 there were 51 absent genes, most of them are from prophage origin, including the strepdornase spd3 gene and a antirepressor. The whole genome alignment, using MGAS5005 as standard, showed that Sp2 is a missing a 30kb region located next to the 1,200,000bp position of the reference genome. This missing fragment has the same position and content as the previously described prophage Φ 5005.2 and Φ 370.3. Another difference is the presence, only in Brazilian samples, of LPXTG-2 gene that produces an extracellular matrix binding protein. Genomics comparison revealed one of the invasive strains that differs from others isolates and from emm 1 reference genomes. In addition, the new invasive strain showed differences in the content of virulence factors compared to other isolated in the same outbreak.

Computer aided identification of a hevein-like antimicrobial peptide of bell pepper leaves for biotechnological use

Patrícia Dias Games, Elói Quintas Gonçalves da Silva, Meire de Oliveira Barbosa, Hebréia Oliveira Almeida-Souza, Patrícia Pereira Fontes, Marcos Jorge Magalhães-Jr, Paulo Roberto Gomes Pereira, Maura Vianna Prates, Glória Regina Franco, Alessandra Faria-Campos, Sérgio Vale Aguiar Campos, Maria Cristina Baracat-Pereira

Universidade Federal de Viçosa, Brazilian Agricultural Research Corporation, Universidade Federal de Minas Gerais

Abstract

Abstract Antimicrobial peptides from plants present mechanisms of action that are different from those of conventional defense agents. They are under-explored but have a potential as commercial antimicrobials. Bell pepper leaves ('Magali R') are discarded after harvesting the fruit and are sources of bioactive peptides. This work reports the isolation by peptidomics tools, and the identification and partially characterization by computational tools of an antimicrobial peptide from bell pepper leaves, and evidences the usefulness of records and the in silico analysis for the study of plant peptides aiming biotechnological uses. Aqueous extracts from leaves were enriched in peptide by salt fractionation and ultrafiltration. An antimicrobial peptide was isolated by tandem chromatographic procedures. Mass spectrometry, automated peptide sequencing and bioinformatics tools were used alternately for identification and partial characterization of the Hevein-like peptide, named HEV-CANN. The computational tools that assisted to the identification of the peptide included BlastP, PSI-Blast, ClustalOmega, PeptideCutter, and ProtParam; conventional protein databases (DB) as Mascot, Protein-DB, GenBank-DB, RefSeq, Swiss-Prot, and UniProtKB; specific for peptides DB as Amper, APD2, CAMP, LAMPs, and PhytAMP; other tools included in ExPASy for Proteomics; The Bioactive Peptide Databases, and The Pepper Genome Database. The HEV-CANN sequence presented 40 amino acid residues, 4,258.8 Da, theoretical pI-value of 8.78, and four disulfide bonds. It was stable, and it has inhibited the growth of phytopathogenic bacteria and a fungus. HEV-CANN presented a chitin-binding domain in their sequence. There was a high identity and a positive alignment of HEV-CANN sequence in various databases, but there was not a complete identity, suggesting that HEV-CANN may be produced by ribosomal synthesis, which is in accordance with its constitutive nature. Computational tools for proteomics and databases are not adjusted for short sequences, which hampered HEV-CANN identification. The adjustment of statistical tests in large databases for proteins is an alternative to promote the significant identification of peptides. The development of specific DB for plant antimicrobial peptides, with information about peptide sequences, functional genomic data, structural motifs and domains of molecules, functional domains, and peptide-biomolecule interactions are valuable and necessary. Supported by FAPEMIG, CNPq, CAPES, and FINEP/Brazil. Thanks to DFP/UFV-Viçosa/MG, EMATER-Coimbra/MG, BIOAGRO/UFV, NuBioMol/UFV and LEM/CENARGEN-Brasília/DF, Brazil.

In silico selection of immunoglobulin sequences produced by phage display technology

Heidi Muniz, Rafael Trindade Burtet, Thaís Costa Lamounier, Tainá Raiol, Nalvo Franco Almeida, Andrea Queiroz Maranhão, Marcelo Macedo Brigido

Universidade de Brasília, Fundação Oswaldo Cruz, Universidade Federal de Mato Grosso do Sul

Abstract

Phage display technology of antibody fragment has diverse applications such as recombinant antibody development, clinical diagnostics, and vaccine research towards a target molecule. Recent improvement in phage display includes an OMICS approach, where individual clones enrichment in a library are traced prior and after ligand selection. Using NGS platforms, phage libraries are retrieved as a large amount of reads, imposing difficulties to analyze data and to predict sequences that potentially bind to target. In this work, we aimed to develop an in silico method to retrieve immunoglobulin sequences from NGS data and estimate enrichment after antigen selection of antibody phage display libraries. The enrichment data are then used for candidate antibody sequence proposition. Two criteria guide our approach. First, a valid immunoglobulin sequence must contain all canonical sequence signatures of the antibody variable domain. Second, a candidate sequence must be enriched along the rounds of affinity selection in phage display experiment. The five steps of the in silico selection are sequence filtering, translation/pattern detection, enrichment analysis, numbering and germline classification. Two sets of NGS sequencing data were tested. The complete analysis may be performed in less than four hours, mostly due to the efficiency of translation and frequency calculation programs. As final output, we have a list of candidate sequences, enriched and identified as immunoglobulin variable domain, and their respective germline classification. Fold change over 700 times could be detected in one of the analysis. Under the proposed criteria, variable domain identification and enrichment estimation, we were able to analyze NGS phage libraries and identify enriched immunoglobulin sequences, that are potential binders of a given target molecule. We believe that the method would be more compatible with different applications of molecular immunology if the complete analysis could be automated. So as future work, we wish to release an automated improved version of this method and bring a friendlier interface.

The identification of DNA binding regions of the $\sigma 54$ factor using artificial neural network

Lucas Martins Ferreira, Roberto Raittz, Jeroniza Nunes Marchaukoski, Vinícius Almir Weiss, Izabella Castilhos Ribeiro dos Santos-Weiss, Paulo Afonso Bracarense, Ricardo Voyceik, Liu Un Rigo

Federal University of Parana

Abstract

Transcription of many bacterial genes is regulated by alternative RNA polymerase sigma factors as the sigma 54 ($\sigma 54$). A single essential σ promotes transcription of thousands of genes and many alternative factors promote transcription of multiple specialized genes required for coping with stress or development. Bacterial genomes have two families of sigma factors, sigma 70 ($\sigma 70$) and sigma 54 ($\sigma 54$). $\sigma 54$ uses a more complex mechanism with specialized enhancers-binding proteins and DNA melting and is well known for its role in regulation of nitrogen metabolism in proteobacteria. The identification of these regulatory elements is the main step to understand the metabolic networks. In this study, we propose a supervised pattern recognition model with neural network to identify Transcription Factor Binding Sites (TFBSs) for $\sigma 54$. This approach is capable of detecting $\sigma 54$ TFBSs with sensitivity higher than 98% in recent published data. False positives are reduced with the addition of ANN and feature extraction, which increase the specificity of the program. We also propose a free, fast and friendly tool for $\sigma 54$ recognition and a $\sigma 54$ related genes database, available for consult. S54Finder can analyze from short DNA sequences to complete genomes and is available online. The software was used to determine $\sigma 54$ TFBSs on the complete bacterial genomes database from NCBI and the result is available for comparison. S54Finder does the identification of $\sigma 54$ regulated genes for a large set of genomes allowing evolutionary and conservation studies of the regulation system between the organisms. The tool and the $\sigma 54$ database are freely available in web respectively in `http://200.236.3.16/s54.php` and `http://200.236.3.66/blast/blast.html`.

Tissue-aware age prediction from DNA methylation data

Marcelo Rodrigo Portela Ferreira, Ricardo Prudêncio, Wolfgang Wagner, Ivan Costa

Universidade Federal da Paraíba, Universidade Federal de Pernambuco

Abstract

Background The human aging is a complex process and the prediction of biological age from biomarkers has important practical applications in many fields such as, forensics, disease treatment and geriatrics. It has been previously observed that changes in DNA methylation were correlated to biological aging and cancer. Linear regression models have been widely used for age prediction from DNA methylation data. Moreover, it is known that the accuracy of age prediction from DNA methylation data are not the same across distinct tissues. Most previous works use curated data to derive DNAm signatures from single tissues. Alternatively multi-tissue predictors were applied for the analysis of large compendia over distinct tissues without taking tissue source information into account. Results In this work, we evaluate four strategies for identification of single tissue or multi-tissue DNA methylation signatures based on penalized regression models. Moreover, we introduce tissue-aware modelling by either including dummy variables representing the tissue types or through a modified Sparse Group Lasso approach which is able to blend tissue specific signatures and non-tissue specific signatures. We evaluate the strategies in a benchmarking data comprising of data sets from 12 different tissue types. For most of the cases, the tissue-aware modelling outperformed the strategies that do not use the tissue information. This was particularly the case in tissues with low number of samples. Our experimental results indicate that take into account the tissue source information leads to an improvement on age prediction accuracy. Avalability Pre-processed benchmark datasets, predictors and source code are available at `http://costalab.org/multiaging`.

Bacterial whole genome comparison: a systematic literature review

Vivian Pereira, Priscilla Wagner, Luciano Digiampietri

University of Sao Paulo (USP)

Abstract

Background: Comparative genomics is useful for the annotation of coding regions, gene function prediction, gene rearrangements and duplications detection and understanding organisms evolution. These knowledge can increase our understanding of bacterial particularities which can lead to the development of vaccines to overcome bacterial diseases, for example. The genome comparison is even more important when complete genomes are compared because of the meaningful and vast characteristics that can be retrieved from the genomes. The aim of this paper is to present the results of a systematic literature review focused on techniques and methods for bacterial whole genome comparison. Results: The results analysis of the 13 primary studies included in the conduction step of the systematic review indicates that genome comparison is performed by the use of data structures such as graphs and suffix trees and by machine learning models such as k-means and hidden Markov models. Some other techniques were also used, for example, common intervals approach and spectogram analysis. Some studies presented better results when the method were applied in more complex genomes than of bacterial genomes. In this case, it is necessary to improve the method to perform better results when comparing bacterial genomes or change the approach to consider the particularities of bacterial genomes. Another aspect to be observed is the studies do not perform a precisely and detailed validation of their methods or techniques. Conclusion: The results of the systematic review presented a state of art overview about the complete bacterial genome comparison and also allowed the identication of researches opportunities based on what was observed in the primary studies.

Comparative genomics analysis uncovers candidate drug targets for Malaria: Using workflows for drug targeting

Kary A. C. S. Ocaña, Daniel de Oliveira, Marco T. A. Garcia-Zapata, Marta Mattoso

National Laboratory for Scientific Computing, Federal Fluminense University, Federal University of Goiás, Federal University of Rio de Janeiro

Abstract

This article presents new cysteine proteases (CPs) found in the genomes of protozoan that pose as potential drug targets in neglected tropical diseases (NTDs). The "discovery" utilizes a very singular approach for drug targeting, absence in host and presence in vector/protozoan (V+H-) and was helped by the systematic execution support of a bioinformatics workflow, named TargFlow. We explored the NTDs and CPs using TargFlow, a genomic-based in silico scientific workflow, which integrates seven databases and five bioinformatics programs. The methodology was broken down into five steps with the steps 1 and 2 collecting the genomes from GOLD, MEROPS and PFAM online databases, step 3 making a comparison using profiles hidden Markov models (pHMMs), step 4 presenting a modified OrthoSelect sub workflow, and finally a manual analysis. Summarizing, it allowed the comparison of 21 genomes against the entire Eukaryotic Orthologous Groups of proteins (KOG) database resulting in the identification of 21 unique CP-related KOGs present in the protozoan and absent in vector/host distributed throughout 15 protozoan genomes. Totaling in an overall presence of 72 CP-related KOGs in 10 diseases (several CPs showing presence in multiple genomes) that pose as possible drug targets, with nine that are completely new and have yet to be explored. Total execution time was 361 hours with only 59 of those being man-hours. The discussion investigated thoroughly the analysis of the results generated for the malaria disease/vector pair and the application of this methodology for other biological components. The bioinformatics workflow TargFlow uncovered nine new cysteine protease drug targets to be investigated in NTDs. This offers a novel approach and use of comparative genomics with clear application to drug targeting across the spectrum of biological components of interest.

A general framework for the gene family-free genome rearrangement problem

Pedro Feijao

Bielefeld University

Abstract

Background: During evolution, genomes are subject to sequence mutation and also large scale rearrangement events. Several distance measures between genomes have been proposed, that can use sequence similarity or genome rearrangements as estimates of evolutionary distance. In the context of structural evolution, a common pre-processing step is to perform an algorithm for gene orthology detection to determine the gene families present in the genomes, in order to represent each gene in every genome as part of a specific family, allowing, for instance, the application of genome rearrangement distance methods. A recent approach called family-free aims to avoid this pre-processing step, receiving as input directly the pairwise similarity between genes. Under this model, the family-free Double-Cut-and-Join distance (FF-DCJ) was proposed. Results: We introduce a family-free framework that can be used with not only the DCJ distance, but any distance defined between genomes (or even strings), also with some improvements over the original FF-DCJ distance, such as adding a contribution for gene insertions and deletions on the distance measure and a way to balance the contributions from the rearrangement distance and sequence dissimilarity to achieve a combined measure of evolution. We show that this problem is APX-hard for several known distances, and give an example of its application using the Single-Cut-or-Join distance, along with an ILP to solve it. Experimental results show how the family-free rearrangement problem can be used as a measure of genome evolution to aid in phylogenetic reconstruction and to estimate the number of rearrangements that occurred during evolution.

Rqc - a Bioconductor package for quality control of high-throughput sequencing data

Welliton Souza, Benilton de Sá Carvalho, Iscia Lopes-Cendes

Universidade Estadual de Campinas

Abstract

As sequencing costs drops with the constant improvements in the field, next-generation sequencing becomes one of the most used technologies in biological research. It allows the detailed characterization of events at the molecular level, including gene expression, genomic sequence variants and well as genomic structural variants, such as copy number variations. Results of such experiments usually yield billions of sequenced nucleotides and each one of them is associated to a quality score. Several software tools allow the assessment of the quality of the whole experiment. However, users often need to switch between software environments to perform all steps of data analysis, adding an extra layer of complexity to the data analysis workflow. We developed Rqc, a Bioconductor package designed to assist the analyst during quality assessment of high-throughput sequencing data. The package is optimized to efficiently process large datasets, regardless their sequencing platforms, by using parallel computing strategies. We created new visualizations of quality data through the use of established analytical procedures, improving the ability of identifying patterns that may affect downstream procedures or technical sources of undesired variation. The software provides a framework for writing customized reports that integrate seamlessly to the R/Bioconductor environment, including publication-ready images. The package also offers an interactive tool to generate quality reports dynamically. Rqc provides a streamlined strategy for quality control assessment of sequencing data. It can be easily added to existing analytical workflows, in particular those created within the Bioconductor ecosystem. It includes capabilities not available through other tools such as customized reports, interactive tool for report generation, fine tuning of parameters, sequencing platform independence, and implementation of parallel strategies to handle large datasets successfully. Rqc is implemented in R and it is freely available through the Bioconductor project (http://bioconductor.org/packages/Rqc) for Windows, Linux and Mac OS X operating systems.

CALI: A novel visual model for frequent pattern mining in protein-ligand graphs

Susana Medina G., Alexandre V. Fassio, Sabrina A. Silveira, Carlos H. da Silveira, Raquel C. de Melo-Minardi

UFMG, UFV, Campus Avançado de Itabira, UNIFEI

Abstract

Protein-ligand interaction (PLI) networks show how proteins interact with small non-protein ligands and can be used to study molecular recognition, which plays an important role in biological systems. The binding and interaction of molecules depend on a combination of conformational and physicochemical complementarity. There are several methods to predict protein-ligand interactions, but a few are designed to identify and describe implications of intelligible factors in protein-ligand recognition. We propose CALI (Complex network-based Analysis of protein-Ligand Interactions), a strategy based on complex network model of protein-ligand interactions revealing frequent and relevant patterns among them. CALI uses a structure superposition followed by a sequence alignment, merging protein atoms. Ligand atoms were also merged considered equivalent according to their physicochemical properties. Patterns obtained with CALI were compared to some determined relevant protein-ligand interactions from precursor experimental studies for Ricin and CDK2 dataset. CALI found all residues of Ricin that interact with ligands. For CDK2 dataset, CALI found 90% of such residues. CALI was able to predict residues experimentally determined as relevant in protein-ligand interaction for both datasets. This new model does not require running data mining algorithms to find the most common protein-ligand interactions since we used network topological properties with a powerful visual and interactive representation of data. CALI is not computationally expensive, avoiding the expensive exact computation of frequent subgraph mining and the subgraph isomorphism to map a frequent subgraph to the input graph. Furthermore, our strategy provides a general view of the input dataset interactions, showing the most common protein-ligand interactions from a global perspective.

Generating transcriptional networks through using text mining techniques for Prokaryotic organisms

Rafael Pereira, Hugo Costa, Sónia Carneiro, Giovani Librelotto, Miguel Rocha, Rui Mendes

University of Minho, SilicoLife, Federal University of Santa Maria

Abstract

Recent advances in systems biology have shown that Transcriptional Regulatory Networks (TRNs) are an important tool to understand cell behaviour. Nowadays, it is possible to find most of the information needed to build TRNs in biomedical literature. Nevertheless, the vast amount of scientific papers makes this process quite difficult to accomplish. Furthermore, there is a large number of databases that store this kind of information and this number has been steadily increasing. These databases do not follow a standard for representing biological information and thus there are several inconsistencies regarding the name of genes, proteins and identifiers. The considerable amount of information implies that the manual curation task is very time consuming, tedious and error prone. Text Mining approaches can help automate this task. In this work we introduce an integrative approach for building TRNs by retrieving relevant information concerning the target organism from both databases and literature and applying text mining techniques, provided by the Note2 framework, in order to extract biological terms and using them to create a dictionary of names and synonyms for genes, proteins, transcription factors. Hence, gathering all necessary information for building these networks. As a case study, two bacteria (Escherichia coli, substrain K-12 MG1665 and Bacillus subtilis substrain 168) were chosen to validate these methods. This work illustrates how it is possible to perform the reconstruction of TRNs for several organisms. In order to accomplish this goal, we developed an integrated approach for the reconstruction of TRNs that retrieves relevant information from important biological databases and stored it into a repository, named KREN. Also, we applied text mining techniques over this integrated repository in order to build TRNs.

Mirnacle: Machine learning with SMOTE and random forest for improving selectivity in pre-miRNA ab initio prediction

Yuri B Marques, Alcione P Oliveira, Ana Tereza R Vasconcelos, Fabio R Cerqueira

Instituto Federal do Norte de Minas, University of Sheffield, Laboratório Nacional de Computação Científica, Universidade Federal de Viçosa

Abstract

Background: MicroRNAs (miRNAs) are key gene expression regulators in plants and animals. Therefore, miRNAs are involved in several biological processes, making the study of these molecules one of the most relevant topics of molecular biology nowadays. However, characterizing miRNAs in vivo is still a complex task. As a consequence, in silico methods have been developed to predict miRNA loci. A common ab initio strategy to find miRNAs in genomic data is to search for sequences that can fold into the typical hairpin structure of miRNA precursors (pre-miRNAs). The current ab initio approaches, however, have selectivity issues, i.e., a high number of false positives is reported, which can lead to laborious and costly attempts to provide biological validation. This study presents an extension of the ab initio method miRNAFold, with the aim of improving selectivity through machine learning techniques, namely, random forest combined with the SMOTE procedure that copes with imbalance datasets. Results: By comparing our method, termed Mirnacle, with other important approaches in the literature, we demonstrate that Mirnacle substantially improves selectivity without compromising sensitivity. For the three datasets used in our experiments, our method achieved at least 97% of sensitivity and could deliver a two-fold, 20-fold, and 4-fold increase in selectivity, respectively, compared with the best results of current computational tools. Conclusions: The extension of miRNAFold by the introduction of machine learning techniques, significantly increases selectivity in pre-miRNA ab initio prediction, which optimally contributes to advanced studies on miRNAs, as the need of biological validations is diminished. Hopefully, new research, such as studies of severe diseases caused by miRNA malfunction, will benefit from this powerful computational tool.

Mitigating the lack of knowledge about long noncoding RNA: Extracting Biological Functions from Biomedical Literature

Yagoub A.I. Adam, Evandro Eduardo Seron Ruiz, Alessandra Alaniz Macedo

Universidade de São Paulo

Abstract

long non-coding RNAs (lncRNAs) play crucial roles in diverse biological processes, such as regulation of gene expression and in shaping 3D nuclear organization, but they are less known in terms of their functions on these processes. Therefore, most of ongoing research efforts are focusing on prediction of their properties and functions. Here we have aimed to extract the biological functions of lncRNAs from biomedical literature. We succeed to extract biologically significant functions from 2,771 single sentences from scientific papers retrieved from PubMed digital library. These sentences have been filtered to eliminate ones without any mention to a biological function, resulting on a set of 1,890 sentences. In order to extract biological functions of lncRNAs, we have prepared two datasets: (I) a dataset consisting sentences in which lncRNAs activate biological functions, and (II) a dataset of sentences where lncRNAs suppress biological functions. Furthermore, we split these two sentences' dataset into 4-gram terms and afterwards we manually extracted meaning full function annotations based on Gene Ontology terms. We have obtained 389 sentences related to activation of biological functions and 22 sentences related to deactivation processes. However simple natural language processing technique could fulfill our purposes, we were able to extract concepts related to biological functions in this kind of RNAs by discovering and augmenting knowledge by means of automatically extracting information methods. In a near future we aim to use more sophisticated methods such as grammatical rules to extract even more functions and to generate a full database for biological functions related to lncRNAs.

Homology modeling provides structural insights into tospovirus nucleoprotein

Rayane Nunes Lima, Muhammad Faheem, João Alexandre Ribeiro Gonçalves Barbosa, Fernando Lucas Melo, Renato Oliveira Resende

Universidade de Brasília

Abstract

Background Tospovirus is a plant-infecting genus within the family Bunyaviridae, which also includes four animal-infecting genera: Hantavirus, Nairovirus, Phlebovirus and Orthobunyavirus. Compared to these members, the structures of Tospovirus proteins are still poorly understood and the lack of structural information prevents detailed insights into the protein interactions. The rapid progress in the understanding of protein folding mechanisms and the advances in the bioinformatics field have provided reliable tools to modeling and predict three dimensional structures for plant viruses proteins. In this study, we performed, by homology modeling, a structural analysis of the nucleoprotein (N) of tospovirus based on the crystal structures of related viral nucleoproteins. Results Here we used the nucleoprotein crystal structure of LACV (La Crosse virus-Orthobunyavirus) as the template to predict a three-dimensional model for the nucleoprotein of the tospovirus GRSV (Groundnut ringspot virus). The resulting in silico model is a monomer composed of two flexible terminal N and C arms and a globular domain with a positively charged groove in which RNA is deeply encompassed. This model allowed to identify the candidate amino acids residues involved in RNA interaction and N-N multimerization and is consistent with site-directed mutagenesis data previously reported. Moreover, most residues predicted to be involved in these interactions are highly conserved among tospoviruses. Conclusions The examination of the proposed GRSV N model has provided a wide knowledge of the tospovirus N multimerization and RNA binding mechanism. In addition, the detailed analysis of this model is of great significance for further in silico mutational studies.

Improving sensitivity in shotgun proteomics using cost sensitive artificial neural networks and a threshold selector algorithm

Fabio R Cerqueira, Adilson M Ricardo, Alcione P Oliveira, Armin Graber, Christian Baumgartner

Universidade Federal de Vicosa, Centro Federal de Educação Tecnológica de Minas Gerais, University of Sheffield, Genoptix, a Novartis company, Graz University of Technology

Abstract

Background This work presents a machine learning strategy to increase sensitivity in mass spectrometry data analysis for peptide/protein identification. Tandem mass spectrometry is a widely used analytical chemistry technique used to identify proteins in complex mixtures, yielding thousands of spectra in a single run which are then interpreted by software. Most of these computer programs use a protein database to match peptide sequences to the observed spectra. The peptide-spectrum matches (PSMs) must also be assessed by computational tools since manual evaluation is not practicable. The target-decoy database strategy is largely used for PSM assessment. However, in general, the method does not account for sensitivity, only for error estimate. Results In a previous study, we proposed the method MUMAL that applies an artificial neural network to effectively generate a model to classify PSMs using decoy hits with increased sensitivity. Nevertheless, the present approach shows that the sensitivity can be further improved with the use of a cost matrix associated with the learning algorithm. We also demonstrate that using a threshold selector algorithm for probability adjustment leads to more coherent probability values assigned to the PSMs. Our new approach, termed MUMAL2, provides a two-fold contribution to shotgun proteomics. First, the increase in the number of correctly interpreted spectra in the peptide level augments the chance of identifying more proteins. Second, the more appropriate PSM probability values that are produced by the threshold selector algorithm impact the protein inference stage performed by programs that take probabilities into account, such as ProteinProphet. Our experiments demonstrated that MUMAL2 provides a higher number of true positives compared with standard methods for PSM evaluation. This new approach reached around 15% of improvement in sensitivity compared to the best current method. Furthermore, the area under the ROC curve obtained was 0.93, demonstrating that the probabilities generated by our model are in fact appropriate. Finally, Venn diagrams comparing MUMAL2 with the best current method show that the number of exclusive peptides found by our method was nearly 4-fold higher, which directly impacts the proteome coverage. Conclusions The inclusion of a cost matrix and a probability threshold selector algorithm to the learning task further improves the target-decoy database analysis for identifying peptides, which optimally contributes to the challenging task of protein level identification, resulting in a powerful computational tool for shotgun proteomics.

GPCRs from Fusarium graminearum detection, modeling and virtual screening - the search for new routes to control Head blight disease

Emmanuel Bresso, Roberto Togawa, Kim Hammond-Kosack Martin Urban, Bernard Maigret, Natalia Martins

EMBRAPA Genetic Resources and Biotechnology, Rothamsted Research, Rothamsted Research, LORIA

Abstract

Fusarium graminearum (FG) is one of the major cereal infecting pathogens causing high economic losses worldwide and resulting in adverse effects on human and animal health. Therefore the development of new fungicides against FG is an important issue to reduce cereal infection and economic impact. In the strategy for developing new fungicides, a critical step is the identification of new targets against which innovative chemicals weapons can be designed. As several G-protein coupled receptors (GPCRs) are implicated in signaling pathways critical for the fungi development and survival, such proteins could be valuable efficient targets to reduce Fusarium growth and therefore to prevent food contamination. In this study, GPCRs were predicted in the FG proteome using a manually curated pipeline dedicated to the identification of GPCRs. Based on several successive filters, the most appropriate GPCR candidate target for developing new fungicides was selected. Searching for new compounds blocking this particular target requires the knowledge of its 3D-structure. As no experimental X-Ray structure of the selected protein was available, a 3D model was built by homology modeling. The model quality and stability was checked by 100 ns of molecular dynamics simulations. Two stable conformations representative of the conformational families of the protein were extracted from the 100 ns simulation and were used for an ensemble docking campaign. The model quality and stability was checked by 100 ns of molecular dynamics simulations previously to the virtual screening step. The virtual screening step comprised the exploration of a chemical library with 11,000 compounds that were docked to the GPCR model. Among these compounds, we selected the ten top-ranked nontoxic molecules proposed to be experimentally tested to validate the in silico simulation.

Identification of ubiquitylation sites through multiobjective genetic algorithm NSGA-II

Paulo Cardoso, Reginaldo Filho, Claudomiro Sales, Regiane Kawasaki, Manoel Lima, Vitor Lima

Universidade Federal do Pará, Federal University of Rio de Janeiro

Abstract

Ubiquitin-proteasome system plays a critical role in regulating a variety of biological processes such as immune response, inflammation and signal transduction.The identification of ubiquitylated proteins sites is fundamental to fully understand the molecular mechanism of the ubiquitin system. However, time-consuming and labor-intensive are some of the drawbacks when using conventional approaches to identify the potential ubiquitylated proteins sites. Machine learning methods appear as an alternative to solve this problem. Methods like Random Forest, K-Nearest Neighbor (KNN) and Genetic Algorithm (GA) can predict ubiquitylation sites through the analysis of fragment sequence properties or the entire proteome. This study uses the Non-dominated Sorting Genetic Algorithm II to identify and predict ubiquitylated sites through concise decision rules. The classification is performed via analysis of feature vectors formed by two types of properties: pysico-chemical properties information retrieved from AAIndex and Amino Acid Composition. NSGA-II builds classification rules based on fragment sequence properties from four organisms-specific datasets containing ubiquitylation sites, namely: S.dataset (S. cerevisiae), H.dataset (H. sapiens), M.dataset (M.musculus) and A.dataset (A.thaliana). Also a general dataset containing all fragments named G.Dataset are generated. Random Forest and KNN are also used to predict ubiquitylation sites in datasets. The best prediction accuracy obtained from independent tests are 68.76%, 84.17%, 69.71%, 71.69%, 76.79% for G.Dataset, S.Dataset, H.Dataset, M.Dataset and A.Dataset respectively. We also generated a list of the most relevant database attributes in classification based on their occurrences in rules. We identified which parameters appear most frequently among the best classification rules, this can provide insights about ubiquitylation sites characteristics. The results also reinforce previous works that state that no universal predictive algorithm exists. In a non-organism-specific dataset is harder to reach good prediction results, although classifiers can reach better results in organism-specific databases.

An Integrative in-silico Approach for Therapeutic Target Identification in the Human Pathogen Corynebacterium diphtheria

Syed Babar Jamal, Syed Shah Hassan, Sandeep Tiwari, Marcus V Viana, Leandro de Jesus Benevides, Asad Ullah Javed Ali Adrián G Turjanski Debmalya Barh, Preetam Gosh, Henrique C P Figueiredo, Artur Silva Vasco AC Azevedo

Universidade Federal de Minas Gerais, Islamia College University Peshawar, Kohat University of Science and Technology Kohat, Universidad de Buenos Aires, Pabellón II, Buenos Aires, Centre for Genomics and Applied Gene Technology, Virginia Commonwealth University, Federal University of Para

Abstract

Corynebacterium diphtheriae (Cd) is a gram-positive human pathogen responsible for diphtheria infection and once regarded for high mortalities worldwide. The fatality gradually decreased with improved living standards and further alleviated when many immunization programs were introduced. However, numerous drug-resistant strains emerged recently that consequently decreased the efficacy of current therapeutics and vaccines, thereby obliging the scientific community to start investigating new therapeutic targets in pathogenic microorganisms. In this context, our contributions include the prediction of modelome of 13 C. diphtheriae strains, using the MHOLline workflow. Considering the quality of the models and using in-house scripts, a set of 465 conserved proteins were selected by combining the results of pangenomics based core- genome and core-modelome analyses. Further, using subtractive proteomics and modelomics approaches for target identification, a set of 23 proteins was selected as essential for the bacteria. Considering human as a host, 8 of these proteins (glpX, nusB, rpsH, hisE, smpB, bioB, DIP1084 and DIP0983) were considered as essential and non-host homologs and were subjected to virtual screening using three different compound libraries (extracted from ZINC database, plant-derived natural compounds and Di-terpenoid Iso-steviol derivatives). The proposed drug molecules showing favorable interactions, lowered energy values and high complementarity with the predicted targets have also been reported in the present study. Our proposed approach expedites the rapid and efficient selection of C. diphtheriae putative proteins for developing a broad-spectrum of novel drugs and vaccines, owing to the fact that some of these targets have already been identified and validated in other organisms.

The Druggable Pocketome of Corynebacterium diphtheriae as a Tool for Novel Targets Identification

Syed Shah Hassan, Leandro G Radusky Syed Babar Jamal Sandeep Tiwari Paulo Vinicius Sanches Daltro de Carvalho, Javed Ali Asad Ullah Henrique C Figueiredo, Debmalya Barh, Artur Silva, Adrian Gustavo Turjanski Vasco AC Azevedo

Universidade Federal de Minas Gerais, Universidad de Buenos Aires, Kohat University of Science and Technology (KUST), Islamia College University Peshawar, Centre for Genomics and Applied Gene Technology, Universidade Federal do Pará, Universidad de Buenos Aires, Pabellón II, Buenos Aires C1428EHA, Argentina

Abstract

Corynebacterium diphtheriae (Cd) is the etiologic agent of an acute, highly infectious, vaccine-preventable and previously endemic diphtheria. The disease gradually alleviated with improved living standards after many immunization programs were introduced. In this context, polypharmacology, reverse vaccinology, comparative and subtractive genomics are the emerging concepts in the field of drug discovery to select appropriate targets. Here, we employed a stratagem along 13 C. diphtheriae completely genomes that utilizes the structural information of the binding sites to characterize the pocketome druggability. We first computed the whole modelome of reference strain NCTC13129; consisting of 13763 open reading frames (ORFs), with 1253 ($\sim$9%) resulted models. The modelled proteins were blasted against other 12 strains (identity $\geq$ 85%, coverage > 80%) and a set of 438 conserved proteins was obtained. Taking the original model as template, we computed the models of all 438 proteins in other 12 strains and a final set of 401 proteins with adequate models was obtained. Protein pockets for this set were computed and only the targets that have highly druggable (HD) pockets (137 proteins) in all strains were retained. Off-targeting homologous information was also obtained by performing a BLASTp (identity 0% and/or no hits) against the human host proteome followed by essentiality analyses, gave a final set of 10 proteins. Finally, this information was merged with the results of another target identification approach for C. diphtheriae to compare the robustness of both strategies. 3 proteins (hisE, phosphoribosyl-ATP pyrophosphatase, glpX, fructose 1,6-bisphosphatase II and rpsH, 30S ribosomal protein S8) were found in common. We hypothesize that our in silico approach will aid in identifying targets with polypharmacological potential against C. diphtheriae and other pathogens and in new drug-discovery pipelines.

A statistical method for the functional classification of gene regulatory networks

Gustavo H. Esteves, Luiz F. L. Reis

Universidade Estadual da Paraíba, Hospital Sírio-Libanês

Abstract

Background: Gene expression data analysis is of great importance for modern molecular biology, given our ability to measure the expression profiles of hundreds of genes and enabling studies rooted in systems biology. Current methods for this type of data analysis are of limited applicability because unknown parameters must be estimated. In this work, we propose a simple statistical model for the functional classification of gene regulatory networks. Results: Here, we present the mathematical construction of a statistical procedure for testing hypotheses regarding gene regulatory network activation under specific biological conditions. The real probability distribution for the test statistic is presented and evaluated by a simple simulation study based on a relevance network. To illustrate the functionality of the proposed methodology, we also present a simple example based on a small hypothetical network and the functional classification of two KEGG networks, both based on gene expression data collected from gastric and esophageal samples. The obtained results are in accordance with the underlying theory for the hypothetical network and both KEGG networks showed significant activation for normal tissues but not for metaplasia and tumors. Conclusions: The statistical procedure presented here is based on one known probabilistic model and yields interesting results for real datasets. In summary, we devised a simple statistical method that can be used to give rise to interesting biological hypotheses based on gene regulatory networks and gene expression data. Availability: This method was implemented in an R package and is available at the BioConductor project website under the name maigesPack.

A systematic comparative evaluation of biclustering techniques

Victor Padilha, Ricardo Campello

University of São Paulo

Abstract

Biclustering techniques are capable of simultaneously clustering rows and columns of a data matrix. These techniques became very popular for the analysis of gene expression data, since a gene could take part of multiple biological pathways which could be active only under specific experimental conditions. Several biclustering algorithms have been developed in the past recent years. In order to provide guidance regarding their choice, a few comparative studies were conducted and reported in the literature. These studies, however, exhibit one or more limitations. First and foremost, the performance of the methods was evaluated through external measures, that, have more recently been shown to have undesirable properties. Furthermore, their analyses was oversimplified, relying mostly on the values achieved by the external measures. Finally, a small number of real datasets was taken into account. We conducted a comparative study involving thirteen algorithms, which were tested on a synthetic data collection as well as over a more representative number of real datasets. With the former, two different experimental scenarios were studied: noise and overlap of biclusters, and the biclusterings were evaluated with more suitable external measures. For the latter, the biclusters found were assessed by gene ontology enrichment. We provide a more insightful analysis of the results, by carefully analyzing the biclusters found by each algorithm. During the experiments with synthetic data four algorithms achieved superior results, two of them on noisy datasets and the other two on datasets with overlap of biclusters. On real data, our results support that five biclustering algorithms stood out.

A graph database approach to reconstruct and visualize metabolic networks

Waldeyr Mendes Cordeiro Silva, Danilo Jose Vilar, Daniel Silva Souza, Marcelo Macedo Brigido, Maria Emilia Machado Telles Walter, Maristela Terto Holanda

UnB

Abstract

Among the challenges of genome-scale reconstruction of metabolic networks is data storage with data modeling that can represent the complexity of systems biology. Meanwhile, recent NoSQL database paradigms have introduced new concepts of scalable storage and data recovery, more specifically, databases based on graphs, which are versatile enough to work with biological data. In this paper, we propose a graph based database approach to solve the problem of genome-scale reconstruction and the visualization of metabolic networks. Based on data of biochemical reactions curated by the International Union of Biochemistry and Molecular Biology (IUBMB), we have created a core database, called Graph Database of Core Pathways. These database was modeled and built using Neo4J, which stores the set of biochemical reactions, related enzymes and others supplementary data. A graphical web interface enables the submission of annotated enzyme sequences of a given organism. Once submitted, these sequences are related to the reactions already contained in the graph database by match operations. From this point, we can infer for a this organism, its biochemical reactions and the metabolic pathways. The Graph Database of Core Pathways organizes a consolidated knowledge of catalytic action of enzymes in a graph and this data are arranged in a flexible way to permit updating, upgrading and manual cure. This database has 5,582 enzymes, 5,709 reactions, almost seven thousand metabolites, almost nine thousand alternative names for enzymes and a few hundred of links and co-factors. A interactive network visualization can be generated using HTML and javascript libraries. A visual image is naturally derived from the Pathway Core Graph implemented on Neo4j. To evaluate the usefulness of our work, we reconstruct the metabolic pathway of Paracoccidioides lutzii (Pb01) through the method previously described. We obtained a metabolic network with 696 enzymatic reactions, enzymes 699 and 1006 compounds for this organism. This work explores the potential of NoSQL databases, specifically the graph Neo4J, and aims to develop alternative computational approaches for the reconstruction of metabolic networks.

GapBlaster – A graphical gap filler for prokaryote genomes

Pablo de Sá, Fábio Miranda, Adonney Veras, Siomar Soares, Kenny Pinheiro, Luís Guimarães, Vasco Azevedo, Artur Silva, Rommel Ramos

Federal University of Pará, Federal University of Triângulo Mineiro, Federal University of Minas Gerais

Abstract

The advent of NGS technologies resulted in an exponential increase in the number of complete genomes available at biological databases. With this advance, many computational tools were developed to specifically analyze this large amount of data in different steps, from processing and quality filtering to gap filling and manual curation. Tools developed for closing gaps are very useful as they result in more accurate genomes, which will influence downstream analyzes as genomic plasticity and comparative genomics. However, the gap filling step is still a challenge for genome assembly that often requires manual intervention. Here, we present GapBlaster, a graphical application for evaluation and closing gaps. GapBlaster was developed in the Java programming language. The software uses contigs obtained in the assembly to perform an alignment against the draft of the genome/scaffold, using BLAST or Mummer, to close gaps. Then, all identified alignments of contigs that extends through the gaps in the draft sequence are presented to the user for further evaluation on the GapBlaster graphical interface. In order to evaluate the GapBlaster, analyses were done using two datasets. All contigs and scaffolds of the datasets were manually evaluated with GapBlaster to close the gaps. The results of GapBlaster were compared to the GapFiller, another software for closing gaps. GapBlaster presents better results when compared to other similar software and has the advantage of having a Graphical interface for manual curation of the gaps. In addition, the GapBlaster program introduce fewer errors in the gap closing step because the user will determine if a gap is filled properly using the graphical interface.

Leveraging High Performance Computing for Bioinformatics: A Methodology that Enables a Reliable Decision-Making

Mariza Ferro, Marisa F. Nicolás, Guadalupe Saji, Antonio R. Mury, Bruno Schulze

National Laboratory of Scientific Computing

Abstract

Background Bioinformatics could greatly benefit from increased computational resources delivered by High Performance Computing. However, the decision-making about which is the best architecture to deliver good performance for a set of Bioinformatics applications is a hard task. The traditional way is finding the architecture with a high theoretical peak of performance, obtained with benchmark tests. But, this is not an assured way for this decision, because each application of Bioinformatics has different computational requirements, which frequently are much different from usual benchmarks. We developed a methodology that assists researchers, even when their specialty is not high performance computing, to define the best computational infrastructure focused on their set of scientific application requirements. The methodology enables to define representative evaluation tests, including a model to define the correct benchmark. Further, a Gain Function allows a reliable decision-making based on the performances of a set of applications and architectures. It is also possible to consider the relative importance between applications and also between cost and performance. Results The methodology was used in a case study for Bioinformatics in which the main objective is to define the best computational infrastructure to the researches. The definition of representative evaluation tests is made for three applications (BLAST, MUMmer, K-means). The experiments are conducted on two parallel architectures and the results highlights that these applications have different computational requirements, which leads to the choice of different computer architectures for each one. The final evaluation of the results is made with the gain function, which delivers the better gain for each architecture considering their costs and performance of all applications. The results pointed out in how the decision- making is difficult and complex, but it is possible to be made with low risk using the methodology described in this work. Conclusions The use of formal methods such as the methodology presented, is useful and relevant for community of Bioinformatics. Decision-making based on the results evaluated by the methodology, with the Gain Function, allows to leverage scientific progress, because it is possible to determine the computational infrastructure which is really necessary and better to boost researches.

SOFTWARE SURVEY FOR BREAST IMAGE PROCESSING

Francisco Adelton Alves-Ribeiro, Miguel de Sousa Freitas, Benedito Borges da Silva, Francisco das Chagas Alves-Lima, Carla Solange Escórcio-Dourado, Fabiane Araújo Sampaio, Luana Mota Martins

Postgraduate Program of the Northeast Network of Biotechnology (RENORBIO), Northeast, Brazil

Abstract

Background: Computer-aided diagnosis (CAD) systems are computational tools that aim to help medical professionals in their diagnostic decisions. Software for processing images has contributed in many ways to medicine, such as, the diagnosis and treatments of diseases such as breast cancer. Breast cancer is one of the main causes of women death all over the world. However, early detection of the disease increases greatly the possibility of cure. Therefore, several types of computer systems based on image processing are being developed by many research groups in order to aid the radiologist in the accuracy of the diagnosis. The aim of this survey was to search for signs of the current state of technology, identifying image processing software in the global marketplace for assistance in medical diagnosis. Results: The outcomes measured were obtained through the searches done in scientific publications repository Web of Science using the Derwent database, including records of deposits of world patents. Data measurement revealed that there were only 107 patent registrations. Further studies and research need to be designed, since the keyword ”image processing software” showed only 3 patent registrations. Conclusions: The United States is the leader in the number of patents per country, with a total of 40 registrations in the abstract search. The title search resulted in only 20 registrations, although this quantity was much larger than that found in other countries. The this survey showed that there is a paucity of patent registration in lesser-developed and developing countries. The reasons may be the absence of a strategic enterprising vision for encouragement to improve technical and scientific endeavors.

BMPOS: The most flexible and user-friendly tool sets for microbiome studies.

Victor Pylro, Daniel Morais, Francislon de Oliveira, Fausto dos Santos, Leandro Lemos, Guilherme Oliveira, Luiz Roesch

Centro de Pesquisa René Rachou - Fiocruz, Universidade de São Paulo - USP, Vale Technology Institute - ITV, Universidade Federal do Pampa - Unipampa

Abstract

Recent advances in science and technology are leading to a revision and re-orientation of methodologies, addressing old and current issues under a new perspective. Advances in Next Generation Sequencing (NGS) are allowing comparative analysis of the abundance and diversity of whole microbial communities, generating a large amount of data and findings at a systems level. The ability to obtain millions (or billions) of microbial sequences from complex samples makes this approach widely used among researchers. The current limitation for biologists has been the increasing demand for computational power and training required for processing of NGS data. Here, we describe the deployment of the Brazilian Microbiome Project Operating System (BMPOS), a flexible and user-friendly Linux distribution dedicated to microbiome studies. The Brazilian Microbiome Project (BMP) has developed data analyses pipelines for metagenomic studies (phylogenetic marker genes), conducted using the two main high-throughput sequencing platforms (Ion Torrent and Illumina MiSeq). The BMPOS is freely available and possesses the entire requirement of bioinformatics packages and databases to perform all the pipelines suggested by the BMP team. The BMPOS may be used as a bootable live USB stick or installed in any computer with at least 1GHz CPU and 512 MB RAM, independent of the operating system previously installed. The BMPOS has proved to be effective for sequences processing, sequences clustering, alignment, taxonomic annotation, statistical analysis and plotting of metagenomic data. The BMPOS has been used during several metagenomic analysis courses, being valuable as a tool for training and an excellent starting point to anyone interested in performing metagenomic studies. The BMPOS and its documentation are available at: `http://www.brmicrobiome.org`.

Optimizations in multiple sequence alignment algorithm using parallel score estimating and ant colony

Geraldo Francisco Donega Zafalon, Evandro Augusto Marucci, Leandro Alves Neves, Carlos Roberto Valencio, Anderson Rici Amorim, Adriano Mauro Cansian, Jose Roberto Almeida Amazonas, Liria Matsumoto Sato, Jose Marcio Machado

University of Sao Paulo, Sao Paulo State University

Abstract

The analysis of biological sequences has become one of the main focus of the biologists. Thus, the computer scientists played an important role, because the analyses of these sequences do not show efficiency without the support of computer programs. Then bioinformatics has been developed. However, due to the increase of the amount of sequences stored in genomic databases, the analyses of these sequences could not be performed only with sequential programs. Thus, parallel computing has added its power to bioinformatics through parallel algorithms to align and analyze the sequence sets. Together with parallel algorithms, some heuristics can be used to reduce the computational costs when the algorithms are executed.Then, this work shows some results obtained by an implementation of a parallel multiple sequence alignment algorithm using score estimating technique and ant colony heuristics. It can be realized that the new approach using the score estimating and ant colony algorithms has better performance than the standard approach. The improvement in the execution time from the new approach to the standard one is around 17% for nucleotides and 14% for aminoacids, which is a significant difference. Moreover, there are results to show that the quality of the final alignments produced by the new approach are slightly better when compared with other tools available and also when it is compared with a standard approach.The results presented here show improvements both in time performance and score quality. The obtained results are relevant because the improvement of the performance with the assurance of alignments make this approach interesting and very useful for computer scientists and biologists.

SIMBA: a web tool for managing bacterial genome assembly

Diego C. B. Mariano, Felipe L. Pereira, Edgar L. Aguiar, Letícia C. Oliveira, Leandro Benevides, Luís C. Guimarães, Edson L. Folador, Thiago J. Sousa, Preetam Ghosh, Debmalya Barh, Henrique C. P. Figueiredo, Artur Silva, Rommel T. J. Ramos, Vasco A. C. Azevedo,

Federal University of Minas Gerais, Virginia Commonwealth University, Centre for Genomics and Applied Gene Technology, Federal University of Pará, Federal University of Pará

Abstract

Background: The evolution of Next-Generation Sequencing (NGS) has considerably reduced the cost per sequenced-base, allowing a significant rise of sequencing projects, mainly in prokaryotes. However, the range of available NGS platforms requires different strategies and software to correctly assemble genomes. Different strategies are necessary to properly complete an assembly project, in addition to the installation or modification of various software. This requires users to have significant expertise in these software and command line scripting experience on Unix platforms, besides possessing the basic expertise on methodologies and techniques for genome assembly. These difficulties often delay the complete genome assembly projects. Results: In order to overcome this, we developed SIMBA (SImple Manager for Bacterial Assemblies), a freely available web-tool that integrates several component tools for assembly and finish bacterial genomes. SIMBA provides a friendly and intuitive user interface so bioinformaticians, even with low computational expertise, can work under a centralized administrative control system of assemblies managed by the assembly center head. SIMBA guides the users to execute assembly process through simple and interactive pages. SIMBA workflow was divided in three modules: (i) projects: allows a general vision of genome sequencing projects, in addition to data quality analysis and data format conversions; (ii) assemblies: allows de novo assemblies with the software Mira, Minia, Newbler and SPAdes, also assembly quality validations using QUAST software; and (iii) curation: presents methods to finishing assemblies through tools for scaffolding contigs and close gaps. We also presented a case study that validated the efficacy of SIMBA to manage bacterial assemblies projects sequenced using Ion Torrent PGM. Conclusion: Besides to be a web-tool for genome assembly, SIMBA is a complete genome assemblies project management system, which can be useful for managing of several projects in laboratories. SIMBA source code is available to download and install in local webservers at `http://ufmg-simba.sourceforge.net`. We also turn available a Linux virtual machine based with SIMBA version 1.2 installed (including demo genome assembly) at the same URL.

SnoReport 2.0: new features and a refined Support Vector Machine improve snoRNA identification

João Victor de Araujo Oliveira, Fabrizio Costa, Rolf Backofen, Peter F. Stadler, Maria Emília M. T. Walter, Jana Hertel

University of Brasilia, Albert-Ludwigs University of Freiburg, University of Leipzig

Abstract

snoReport was proposed to identify the two main classes of snoRNAs (box H/ACA and box C/D), using secondary structure prediction combined with machine learning. In this work, we present snoReport 2.0, a significant improvement in the original method: extracting new features for both box C/D and H/ACA box snoRNAs; developing a more sophisticated technique in the SVM training phase with recent data from vertebrate organisms and a careful choice of the SVM parameters C andγ; and using new versions of tools and databases that were taken first to build snoReport. To validate this new version, we tested snoReport 2.0 in different organisms. These experiments showed a very good performance. Results of the training and test phases of boxes H/ACA and C/D snoRNAs, in both versions of snoReport, are discussed. Validation on real data was performed to evaluate the predictions of snoReport 2.0. Our program was applied to a set of previously annotated sequences, some of them experimentally confirmed, of humans, nematodes, drosophilids, platypus, chickens and leishmania. We significantly improved the predictions for vertebrates, since the training phase used information of these organisms, but H/ACA box snoRNAs identification was improved for the other ones. We presented snoReport 2.0, to predict H/ACA box and C/D box snoRNas, an efficient method to find true positives and avoid false positives in vertebrate organisms. H/ACA box snoRNA classifier showed a F-score of 93% (an improvement of 10% regarding the previous version), while C/D box snoRNA classifier, a F-Score of 94% (improvement of 14%). Besides, both classifiers exhibited performance measures above 90%. These results show that snoReport 2.0 avoid false positives and false negatives, allowing to predict snoRNAs with high quality. In the validation phase, snoReport 2.0 predicted 67.43% of vertebrate organisms for both classes. For Nematodes and Drosophilids, 69% and 76.67%, for C/D box and H/ACA box snoRNAs were predicted, respectively, showing that snoReport 2 is good to identify snoRNAs in vertebrates and also H/ACA box snoRNAs in invertebrates organisms.

Towards the Semantic Composition of Gene Expression Analysis Services

Gabriela Der Agopian Guardia, Luís Ferreira Pires, Eduardo Gonçalves da Silva, Cléver Ricardo Guareis de Farias

Universidade de São Paulo, University of Twente

Abstract

Background: Gene expression studies often require the integrated use of a number of analysis tools. Manual integration of gene expression analysis tools can be cumbersome and error prone. In order to support a higher level of automation in the integration process, efforts have been made towards the provision of analysis tools as semantic web services. In addition, different software environments have been defined to support semantic service composition in the biomedical domain. However, most of these environments require users to focus on technical details of the composition process. Moreover, the proposed approaches consider only the execution of simple services behaviours. Results: In this paper, we propose a novel approach for the semantic composition of gene expression analysis services that addresses the shortcomings of existing approaches. To position our approach, we first identify the requirements of gene expression analysis and propose a composition life-cycle that captures these requirements. We then define a layered architecture to support the proposed life-cycle. Our architecture is designed to provide a higher level of abstraction, thus enabling users (biologists) to focus on biological research questions. In addition, it provides support for the definition of complex service behaviours. We conclude by providing an overview of the supporting infrastructure we have built to implement and validate our architecture. Conclusions: The proposed approach provides a high level of abstraction thus allowing biologists to focus on the biological studies being performed rather than on technical details of the composition process. Moreover, our supporting infrastructure addresses all phases of the composition life-cycle and includes a prototype implementation of a graphical interface to support end users during the creation and execution of analysis workflows.

SwiftGECKO: a provenance-enabled parallel comparative genomics workflow

Maria Luiza Mondelli, Oscar Torreño, Kary A. C. S. Ocaña, Marta Mattoso, Michael Wilde, Ana Tereza Vasconcellos, Oswaldo Trelles, Luiz M. R. Gadelha

National Laboratory for Scientific Computing, Federal University of Rio de Janeiro

Abstract

Biology has moved from gene-gene to overall genomes analysis, both somehow pushed up for improvements in DNA sequencing technology and the availability of high performance computing (HPC) resources. Conducting computer-based comparative genomic experiments is a complex and time-consuming task since a huge quantity of data needs to be processed and a large set of programs are used as the income of another. This coherent flow can be designed as scientific workflows. Scientists may require technologies as parallel scientific workflow management (SWfMS) and HPC environments for improving the total processing time and assisting scientists at the management, treatment and analyses of the data. We propose the workflow SwiftGECKO, an updated version of the sequential application GECKO for genome comparisons based on the fast identification of high-scoring segment pairs (HSPs). SwiftGECKO was implemented in the Swift parallel functional dataflow system, providing benefits such as the intrinsic parallelism for execution and provenance data management. We tested SwiftGECKO using a dataset of 40 evolutionary related bacterial genomes. SwiftGECKO keeps a detailed domain data provenance trace of the experiment in a relational database. General benefits regard the capacity to retrieve domain data associated with computational ones. In the specific field of comparative genomics, we explain these benefits with a set of queries aimed to (i) exploring the biological information contained in resulting files, no prone to error manual manipulation is required from scientists; (ii) tracing the taxonomic lineage and inferring evolutionary relationships of genomes based on annotations of domain data provenance of the experiment (e.g. statistics of hits and fragments); and (iii) driving new task scheduling strategies for the execution of experiments, e.g. estimating a priori the demanded CPU/time for processing other genomes. Regarding computational results, we present performance improvements of up to 89.40% (128 cores) in the execution time when compared to its sequential execution (1 core), which drops from around 2 hours and 50 minutes to 18 minutes using a shared-memory computer with 128 processing cores. The process of multiple comparisons of genomes is considered as a time-consuming and costly experiment. In this article, we have addressed the problem of tracing and managing the provenance data and results information (both domain specific and general computational annotations) of the experiment. Additionally, we propose a parallel solution which significantly reduces the execution time of the sequential executions. The presented results demonstrate that SwiftGECKO is computationally-efficient for parallelizing massive tasks for complete genome comparisons.

Compromise or optimize? The breakpoint anti-median

Caroline Anne Larlee, Alex Brandts, David Sankoff

University of Ottawa

Abstract

The median of k $\geq$ 3 genomes was originally defined to find a compromise genome indicative of a common ancestor. However, in gene order comparisons, the usual definitions based on minimizing the sum of distances to the input genomes lead to degenerate medians reflecting only one of the input genomes. "Near-medians", consisting of equal samples of gene adjacencies from all the input genomes, were designed to restore the idea of compromise to the median problem. We explore adjacency sampling constructions in full generality in the case k = 3, with given overlapping sets of adjacencies in the three genomes, where all adjacencies in two-way or three-way overlaps are included in the sample. We require the construction to be maximal, in the sense that no additional proportion of adjacencies from any of genomes may be added without violating the local linearity of the genome. We discover that in incorporating as many adjacencies as possible, evenly from all the input genomes, we are actually maximizing, rather than minimizing, the sum of distances over all other maximal sampling schemes. We propose to explore compromise instead of parsimony as the organizing principle for the small phylogeny problem. The anti-median genomes are constructed to have precise normalized distances from g1, g2, and g3, in the sense of their limiting behaviour as n $\longrightarrow \infty$. This behaviour is predicated on the inclusion of all the adjacencies in the two-way and three-way overlap, and the completion of the sampled genome by random matching of unpaired ends. Outside of this class are arbitrary random genomes whose scores all approach 3, and those completed by maximum matching algorithms, whose scores are less than those of randomly completed samples.

A systematic review of phylogenetic analysis of a specific gene

Priscilla Koch Wagner, Vivian MY Pereira, Luciano Antonio Digiampietri⟶ ∞

University of Sao Paulo (USP)

Abstract

Background: Phylogenetic analysis aims to analyze the evolutionary relationships among different species. It is typically made by comparing the genes of a specie with the genes of other species, looking for similarities. The phylogenetic analysis process involves some complex steps that need computational tools to be realized and integrated. Nowadays, phylogenetic analysis researchers uses some computational tools that are not integrated. So, the communication between them is made manually, being really expensive. Based on that, this paper aims to identify and analyze methods and techniques that have been used on literature for the achievement of the phylogenetic analysis process or part of it, through a systematic review. Results: The results analysis of the studies selected during the systematic review's conduction indicates that phylogenetic analysis commonly uses traditional bioinformatics tools. It was also verified that a few initiatives have proposed different approaches from the traditional for phylogenetic analysis, such as methods based on artificial intelligence or mathematical models. The evaluation of the methods proposed on these papers, in general, were more rigorous than the papers that used traditional tools. Lastly, few approaches propose improvements and automatic integration of tools that implement phylogenetic analysis process steps. Conclusion: The systematic review provided the state of the art overview of the phylogenetic analysis area. The results obtained showed that there are some traditional consolidated bioinformatics tools that are broadly used in the phylogenetic analysis. Despite tools that are consolidated, it is noticed that efforts for their improvement are missing. Neither the integration of phylogenetic analysis process steps is explored in the papers reviewed. Because of the importance of phylogenetic analysis for several areas, it is expected that this review contributes for development of future work that overcome the limitations observed on selected studies, specially related to improvements and integration of consolidated tools.

A Relational Database Representation for Biological Sequences

Sérgio Lifschitz, Edward Hermann Haeusler, Cristian Tristão, Paulo Cavalcanti Gomes Ferreira, Maristela Holanda Maria Emilia Walter

PUC-Rio, UFRJ, UNB

Abstract

Research in molecular biology data modeling generates a large amount of data and they need to be well organized, structured and persisted. Most biological data are stored into text-based files. For large volumes of data, the natural way would be to use a DBMS to manage them. However, these systems do not have adequate structures to represent and manipulate data specific to the biological domain. We usually represent biological sequences as simple strings: text, varchar or BLOBs - binary large objects – data types. This is not only syntactically restricted but, also, there is no way to assign a comprehensive semantics. We need a whole set of compositional, positional and content information that could be considered, in order to ensure an adequate object representation. The management of data (structure, storage and data access) has become a major problem for the bioinformatics research area. We propose a conceptual model for representing biological information that includes the central dogma of molecular biology as well as an ADT - Abstract Data Types - specific for the manipulation of biological sequences and its derivatives. Our approach allows the expression of some particular functions applied to biological sequences that use our proposed ADT. We could mention simple functions such as DNA checking, DNA complement and reverse, as well as more complex ones, like transcription, translations and ORF searching. We contemplate procedures that are even more sophisticated. For example, the identification of unique genes, protein taxonomy and homologous genes. We have tested our approach with actual biological data, including GeneOntology, NCBI Reference Sequence (RefSeq), PFAN and KEGG. The main goal of this work is to improve the quality and interpretation of biological data, besides our understanding of biological systems and their interactions. The database system design is a critical step but the actual representation into known data models remains ad-hoc. This paper describes a way to represent biological sequences and extract relevant information, from simple to complex queries including recursive ones.

AutoModel: new tool for interactive protein homology modeling

Joao Luiz de Almeida Filho, Jorge Hernandez Fernandez

State University of North Fluminense (UENF)

Abstract

The protein tertiary structure prediction is not a simple task but the assessment of this information becomes essential for functional annotation. Computer prediction of protein structure is an important tool in structural biology helping to construct large quantity of interaction model of protein-ligand complexes or used to obtain three-dimensional structure and functional information of non-crystalizing proteins. However, the complexity of modeling softwares and a hard-to-use user interface makes difficult the use for non-expert scientists. On this context was developed a semi-automatic client-server software for protein homology modeling, the AutoModel. The main goal of AutoModel is to provide a graphical and practical interface to perform modeling experiments in a distributed architecture. Our system facilitates the interaction of new users in a full modeling pipeline as follow: 1- Searching structural templates; 2- Alignments of sequences; 3- Protein modelling; 4- Model refinement and 5- Loops refinement. In AutoModel 0.5 development we evaluated the use of different alignment tools in order to increase the quality of generated models, reduce the computational cost, and evaluate the impact of these changes in modeling quality and speed of the experimentation. Modeller alignment algorithm of 9v4 and 9v9 versions, HMMER and Muscle programs were used with the same protein set. Thus, we evaluated: i) the quality of the generated models using the Prosa-web Server; ii) machine time for full modeling experiment and iii) the time lapse of the alignment step. Our data suggest that the use of AutoModel with the 9v9 Modeller version obtained better results than generated models by Modeller 9v4 and the HMMER versions of the program pipeline, demonstrating a dramatic improvement in alignment algorithm of 9v9 version. However, using Muscle as alignment tool in the pipeline increase the quality of obtained models if compared to the other tested versions and obtained significantly lower computational costs, which is always interesting in a distributed system running on a central server as AutoModel. AutoModel is free available for academic community ”as it” in `https://sites.google.com/site/uenfautomodel`.

Comparing genomes with duplicate genes by DCJ and single gene indels

Diego Rubert, Pedro Feijão, Marília Braga, Jens Stoye, Fábio Martinez

Universidade Federal de Mato Grosso do Sul, Bielefeld University

Abstract

Rearrangements are large-scale mutations in genomes, responsible for complex changes and structural variations. Most rearrangements that modify the organization of a genome can be represented by the double cut and join (DCJ) operation. In addition, there are mutations that modify the content of a genome, such as insertions and deletions, jointly called indels. Indels can be restricted so that only a single gene is deleted or inserted at once. In this case they are called single gene indels. Given two genomes, we are interested in the problem of computing the rearrangement distance between them, i.e., finding the minimum number of DCJ operations and single gene indels that transform one genome into the other. We show that, when the genomes have duplicate genes, this problem is NP-hard. Recently, Shao and Lin (BMC Bioinformatics 13, Suppl 19, S13, 2012) studied a related problem that aims to find a number of DCJ operations and single gene indels, that is a lower bound to the rearrangement distance. Their problem is also NP-hard for genomes with duplicate genes, and they proposed a polynomial time algorithm which they claim to be a $(1.5 + \epsilon)$-approximation for their problem. However, we found an infinite family of counterexamples which unfortunately invalidates their claim. Still we could adapt some steps of their approach and develop two new heuristics for solving the problem of computing the rearrangement distance using DCJ and single gene indels. Those heuristics fix an issue in the Shao and Lin's algorithm. Finding an approximation algorithm remains an open problem.