BiP-Day 2014
Seconda Giornata della Bioinformatica Pugliese

BARI, 19 DICEMBRE 2014
POLITECNICO DI BARI
AULA MAGNA “E. ORABONA”
ORGANIZZAZIONE

Coordinatori
D’ELIA Domenica, CNR ITB
LIUNI Sabino, CNR ITB

Comitato Scientifico
D’ELIA Domenica, CNR ITB
GISEL Andreas, CNR ITB
GRILLO Giorgio, CNR ITB
LICCIULLI Vito Flavio, CNR ITB
LIUNI Sabino, CNR ITB
VICARIO Saverio, CNR ITB
PESOLE Graziano, CNR IBBE & UNIBA
CATALANO Domenico, CNR IBBR
ANCONA Nicola, CNR ISSIA
ATTIMONELLI Marcella, UNIBA
DE GRASSI Anna, UNIBA
PICARDI Ernesto, UNIBA
BELLOTTI Roberto, POLIBA & UNIBA
MAGGI Giorgio, POLIBA & UNIBA

Sviluppo & Comunicazione Web
CONSIGLIO Arianna, CNR ITB
D’ELIA Domenica, CNR ITB
LOSITO Nicola, CNR ITB

Segreteria
CANANZI Giuseppe, CNR ITB
PICCOLO Michele, CNR ITB
PROGRAMMA

08:30 – 08:45  Registrazione

08:45 – 09:00  Benvenuto e introduzione ai lavori della Giornata

D’elia Domenica, Istituto di Tecnologie Biomediche, Bari

Sessione I: Le grandi infrastrutture per la Bioinformatica operanti nella Regione Puglia

09:00 – 09:15  Le grandi infrastrutture di ricerca europee: i nodi italiani di Elixir e Lifewatch

Graziano Pesole, Istituto di Biomembrane e Bioenergetica; e Dipartimento di Bioscienze, Biotecnologie e Biofarmaceutica, Università di Bari

09:15 – 09:30  I progetti PON ReCaS e PRISMA: le possibili ricadute sulla bioinformatica pugliese

Giacinto Donvito, Istituto Nazionale di Fisica Nucleare, Sezione di Bari

09:30 – 09:50  Science applications in the Bc2S farm

Pasquale Notarangelo, Istituto Nazionale di Fisica Nucleare, Sezione di Bari

Sessione II: Nuovi Strumenti Bioinformatici e Applicazioni per la Genomica Strutturale e Funzionale

09:50 – 10:10  BioMaS: a friendly web service for improved taxonomic assessment of microbiomes

Bruno Fosso, Istituto di Biomembrane e Bioenergetica, CNR, Bari

10:10 – 10:30  Phylogenetic diversity as an approach to look at change in hierarchical structured data as biological communities

Saverio Vicario, Istituto di Tecnologie Biomediche, Bari

10:30 – 11:00  Coffee break
Results

We tested phyloH to characterize microbiome in the honeybee (Apis mellifera, Insecta: Hymenoptera) and its parasitic mite varroa (Varroa destructor, Arachnida: Parasitiformes). The rationale is that the comparative analysis of honeybee and varroa microbiomes could open new perspectives concerning the role of the parasites on honeybee colonies health. Our results showed a dramatic change of the honeybee microbiome when varroa occurs, suggesting that this parasite is able to influence host microbiome. Among the different approaches used, only the entropy method, in conjunction with phylogenetic constraint as implemented in phyloH, was able to discriminate varroa microbiome from that of parasitized honeybees.

Conclusions

The phylogenetic diversity approach seems more sensitive than distance method and allow a more fine grain data exploration by inspecting the branch contribution to beta diversity. The method need to be enlarged to take in account phylogenetic uncertainty.

References

Marcon Eric, Herault Bruno (2014) https://hal-agroparistech.archives-ouvertes.fr/hal-00946177v1

14:40 – 15:00  nc-aReNA: an integrated web-based platform for non-coding RNA-seq data classification and annotation
Angelica Tulipano, CNR – Institute for Biomedical Technologies, Bari

15:00 – 15:20  Managing NGS differential expression uncertainty with fuzzy sets
Arianna Consiglio, CNR, Institute for Biomedical Technologies and Dept. of Computer Science, University of Bari Aldo Moro, Bari

15:20 – 15:40  Docking-based virtual screening of chemical libraries for the identification of new apoptosis-inducer drugs
Ciro L. Pierri, Department of Biosciences, Biotechnology and Biopharmaceutics, University of Bari, Italy

15:40 – 17:00  Sessione III

Tavola Rotonda su: “Bioinformatica, Formazione, Internazionalizzazione, Rientro dei cervelli e inserimento nei mercati produttivi”

Phylogenetic diversity as an approach to look at change in hierarchical structured data as biological communities

A Sandionigi(1), S Vicario(2), EM Prosdocimi(1), A Galimberti(1), A Ferri(1), A Bruno(1), B Balech(4), V Mezzasalma(3) and M Casiraghi(1)

(1) ZooPlantLab, Department of Biotechnology and Biosciences, University of Milan-Bicocca, Piazza della Scienze 2, 20126 Milan, Italy,
(2) Institute of Biomedical and Technologies (ITB), National Research Council (CNR), Via Giovanni Amendola, 122/D, 70126 Bari, Italy,
(3) DEFENS, University of Milan, Via Mangiagalli, 25, 20133 Milan, Italy
(4) Institute of Biomembrane and Bioenergetics (IBBE), National Research Council (CNR), Via Giovanni Amendola, 165/A, 70126 Bari, Italy

saverio.vicario@ba.itb.cnr.it

Motivations

Biological diversity was early identified in ecological studies as an important emergent property of natural systems, linked to the ability to persist through time. Unfortunately the difficulties in identifying common metrics for diversity contributed to produce a vast literature with unclear conclusion. Within an information theoretical framework it is possible to define diversity as the exponential of Shannon entropy and thereby obtain a self consistent measure of overall within and between group diversity. This framework was recently enlarged to take into account that biological entities (lineage, guild, species, or group of individuals) are differently diverse by using tree shaped similarity structure, allowing to define phylogenetic diversity within the frame of information theory (Chao A, Chiu CH, Jost 2010; Marcon and Herault 2014: Sandionigi A et al 2014).

Methods

21 pairs of larva bees and its parasite were sampled across 7 beehives in northern Italy. The microbiomes were extracted from each biological sample and V3 SSU region was amplified and sequenced with 454 FLX technologies. The reads were de-noised using ACACIA and the chimera detector of USEARCH. We compared the phylogenetic diversity approach as implemented in our tool PhyLOH with the distance approach (Jaccard, Chord, wUnifrac, Unifrac) implemented in QIIME taking in account or not phylogenetic information and abundance.
Conclusions

The protocol was tested on mtDNA samples of colorectal adenocarcinoma from High Throughput and ovarian cancer from Sanger sequencing, enabling in the latter case the inference of also somatic non-synonymous mutations on the strength of prioritization criteria.

References


An experimental solution for accurate screening and interpretation of structural genetic variants

P. Cazzati(1), R. De Molfetta(1), R. Greco(1), L. Loberafo(1), F. Nigro(1), P. Noviello(1), N. Tempesta(1), M. Carella(2), S. Castellana(3), T. Mazza (3), O. Palumbo(2), P. Palumbo(2)

(1) Innovation Lab, Competence Team Telemedicine & Clinical Genomics, Exprivia S.p.A., Molfetta (BA) - Italy
(2) Medical Genetics unit, Casa Sollievo della Sofferenza, San Giovanni Rotondo (FG) - Italy
(3) Bioinformatic Unit, Casa Sollievo della Sofferenza, San Giovanni Rotondo (FG) - Italy

pietro.noviello@exprivia.it
m.carella@operapadrepio.it

Motivations

The progressive cost reduction of the DNA investigation technologies facilitates the acquisition and the employment of genomic data for clinical purposes. Genetic information can drive both clinicians and researchers towards a more personalized medicine. In order to support genome-based diagnosis, Exprivia, as industrial partner for the PON02_00619_3461281 Project in collaboration with Casa Sollievo della Sofferenza as scientific partner, proposed to establish a network between clinical facilities and molecular genetics laboratories.

Methods

The following prototypes software packages have been developed:

- A portal to support the collaboration between physician and laboratories through clinical-genomic data exchange, by collecting clinical data (focused on Human Phenotype Ontology); the opportunity to launch and follow the experimental procedure from the committed laboratory; the acquisition of the corresponding medical report.
- A knowledge base to store the detected genetic alterations and find out possible correlations among phenotypic abnormalities and other clinical data in order to support epidemiologic studies.
- A bioinformatic pipeline for the interpretation of experimental outcomes based on centralized public and local genetic data resources and for the evaluation of the clinical role of the genetic alterations and relative pathogenicity.
Results

We have tested the prototypes software packages for the interpretation of Copy Number Variants (CNVs) coming from array Comparative Genomic Hybridization experiments (‘aCGH’).

The pipeline reduces the time of analysis and facilitates the data interpretation. The portal simplifies the data entry process, thanks to the integration of standardized description of the individual phenotype, and supports the diagnosis evaluation.

Conclusions

This work has led to the development of software modules that manage different data typologies, according to a standard procedure that integrates clinical and genomic information, and correlates them with phenotypic description. From the clinical side, this resource allows the collection of patients anamnesis, its follow-up and the possibility to contribute to epidemiological studies. From the laboratory side, the portal allows to detect novel CNVs, on the basis that shared CNVs from patient with different phenotypes are probably functionally neutral while shared CNVs from patient with similar phenotypes can putatively be considered as “pathogenic” and concur to define new syndromes.

A bioinformatics approach for the prioritization of disease candidate human mtDNA mutations

(1)Department of Sciences and Technologies, University of Sannio, Benevento, Italy
(2)Department of Medical and Surgical Sciences, University of Bologna, Italy
(3)Department of Biosciences, Biotechnologies and Biopharmaceutics, University of Bari, Italy
(4)Department of Molecular Biotechnology and Health Science, University of Torino

ma.santorsola@gmail.com

Motivations

Assigning a pathogenic role to mitochondrial mutations and unveiling the potential involvement of the mitochondrial genome in diseases are challenging tasks in the field of human mitochondrial medicine. With this aim in mind, we designed a prioritization process in order to recognize rare disease-causing mutations for further investigations.

Methods

The prioritization process here implemented is based on the application of MToolBox software [1] to mtDNA sequences. The workflow relies on an exhaustive functional annotation of mtDNA variants such as nucleotide variability, observed in the genomes of healthy subjects stored in HmtDB [2]; annotations from Mitomap; estimation of a disease score based on predictions of pathogenicity for non-synonymous changes.

Results

The disease score was estimated as the weighted mean of the probabilities that an amino acid substitution may be deleterious provided by several pathogenicity predictors. Disease score and nucleotide variability cutoff points were then established in order to discriminate events contributing to a defective phenotype. The prioritization criteria took into account the non-synonymous mutations recognized by rCRS, RSRS and Macro-Haplogroup Consensus Sequences at same time, non-haplogroup defining, featuring nucleotide variability lower than the variability cutoff and disease score higher than disease score cutoff.
1.8 μM respectively, whereas the inhibition constants of the 4 newly identified inhibitors range between 0.2 and 2 μM.

Conclusions

Our analyses and simulations suggest that: 1) our computational strategy successfully identifies molecules with high affinity for AAC 2) in vitro assays show that those molecules are active as AAC inhibitors in the submicromolar range, and thus they can be used as scaffolds for drawing new potential drugs to be emplyed as apoptosis inducers.

References


A Comprehensive Characterization of Mitochondrial DNA Mutations in Glioblastoma Multiforme

Motivations

Glioblastoma Multiforme (GBM) is the most malignant brain cancer in adults, with a poor prognosis, whose molecular stratification still represents a challenge in pathology and clinics. On the other hand, mitochondrial DNA (mtDNA) mutations have been found in most tumors as modifiers of the bioenergetics state, albeit in GBM a characterization of the mtDNA status is lacking to date. Here, a large characterization of the burden of mtDNA mutations in GBM samples was performed.

Methods

First, investigation of tumor-specific vs. non tumor-specific mutations was carried out with the MToolBox bioinformatics pipeline by analyzing 46 matched tumor/blood samples, from whole genome or whole exome sequencing datasets obtained from The Cancer Genome Atlas (TCGA) consortium. Additionally, the entire mtDNA sequence was obtained in a dataset of 104 fresh-frozen GBM samples.

Results

Mitochondrial mutations with potential pathogenic interest were prioritized based on heteroplasmic fraction, nucleotide variability, and in silico prediction of pathogenicity.
A preliminary biochemical analysis of the activity of mitochondrial respiratory complexes was also performed on fresh-frozen GBM samples.

Conclusions

Although a high number of mutations were detected, we report that the large majority of them do not pass the prioritization filters. Therefore, a relatively limited burden of pathogenic mutations is indeed carried by GBM, which did not appear to determine a general impairment of the respiratory chain.

Docking-based virtual screening of chemical libraries for the identification of new apoptosis-inducer drugs

G Punzì, A Onofrio, S Todisco, MA DiNoia, G Parisi, G Redavid, MT Ricatti, A De Grassi, CL Pierri

Department of Biosciences, Biotechnology and Biopharmaceutics, University of Bari, Italy
ciroleopierri@gmail.com / ciro.pierri@uniba.it

Motivations

Most of current cancer chemotherapeutics cause cell death via apoptosis. A recently proposed new powerful chemotherapy target is the mitochondrial ADP/ATP carrier (AAC) [1]. AAC plays a key role in the cell viability because it translocates the synthesized ATP from the mitochondrial matrix to the cytosol. Furthermore, AAC is a key component of the mitochondrial permeability transition pore that regulates mitochondrial apoptosis. Carboxyatractylloside (CATR) and Bongkrekic acid (BKA) are powerful and highly specific AAC inhibitors, and they can induce mitochondrial dysfunction at molecular level and poisoning at physiological level. For identifying other specific AAC inhibitors we performed docking-based virtual screening of an in-house developed chemical library.

Methods

In our analyses we used the crystallized AAC-CATR complex and the Autodock suite [2] for screening a chemical library consisting of 10,000 bioactive compounds (that include also CATR and BKA as reference inhibitors) to identify a set of molecules with a high binding affinity for mitochondrial AAC. Autodock analyses are based on a Free Energy scoring function, already validated on various sets of protein–ligand complex structures [2], which evaluates the fitness between the AAC and the screened ligands. In vitro assays were performed for validating in silico predictions.

Results

From the virtual screening we identified 15 ligands predicted to have similar physical-chemical properties (in terms of binding energy) to those shown by CATR and BKA. The 15 candidates were used in in vitro transport assays and it was found that 4 of them were good AAC inhibitors. The inhibition constants of the known AAC inhibitors, i.e. CATR and BKA, were found to be 7nM and
Million), fold changes and False Discovery Rate (FDR) corrected p-values calculation and differential expression analysis of all (or user-chosen) ncRNA classes, by comparing two or more experimental conditions or time-courses data. An additional module, called IsomiR identification, provides the classification and quantification of miRNA variants, called IsomiRs, identified in the experiments. The pipeline has been designed and implemented using Pentaho Data Integration tool (pdi) and the analysis results and annotations are stored in a data-warehouse implemented using Infobright (http://www.infobright.org). A user-friendly web-based Graphical User Interface (GUI), developed by using the JAVA platform, guides the user in the submission process and displays results in tables and graphs.

Results

The main features of the nc-aReNA are:

- identification and classification of reads in known functional ncRNA classes and dataset export
- identification and filtering of reads mapping to ribosomal RNAs and mtDNA transcripts;
- quantification of ncRNA expression and differential expression analysis
- graphical visualization of sample expression profiles in different conditions and at different time courses
- additional annotation such as target genes and pathways for miRNAs;
- creation of a collection of unclassified reads, useful for the prediction of novel ncRNAs

As test case, we have used Illumina small RNA-seq data produced for the expression profiling of smallRNAs in different tissues isolated from M. musculus. This experiment includes three time-courses and three technical replicates from two different tissues.

Conclusions

nc-aReNA is a web-based platform which provides a user-friendly GUI for the analysis and classification of human and mouse small RNA-Seq datasets. It has been developed to support biologists and clinicians, with no prior specific computer science knowledge, in the annotation and biological interpretation of NGS data. The platform is designed for to be used with any type of organism provided that basic information is available (i.e. reference genome and ncRNAs).

Managing NGS differential expression uncertainty with fuzzy sets

Consiglio A(1,2), Mencar C(2), Grillo G(1), Liuni S(1)
(1) CNR, Institute for Biomedical Technologies, Bari, Italy
(2) Dept. of Computer Science, University of Bari Aldo Moro, Bari, Italy
arianna.consiglio@ba.itb.cnr.it

Motivations

One of the major issues of Bioinformatics is trying to translate heterogeneous and fuzzy biological concepts into representations that can be processed by computers. In this work we propose a workflow that manages the uncertainty arising throughout the analysis of differential expression in case-control NGS data comparison. Such uncertainty is described and exploited in the computation through the use of fuzzy sets.

Methods

The aim of this work is to obtain a list of candidate differential expression (DE) events, ordered by importance, like every DE analysis tool does. The importance criteria take in account some uncertainty issues typical of DE analysis. First of all, we deal with uncertainty in read mapping results. When a read maps to more than one reference sequence it is called multiread. When such event is sporadic, many researchers choose to exclude multireads from the analysis, but this option leads to an underestimation of the read counts. Another approach is to estimate the real number of read counts using probabilistic models, based on some assumptions on the distribution of data. We model such uncertainty with fuzzy sets describing the possibility that each gene has a given read count. After this step, we analyze the trend of the Fold Change (FC) on varying the order of magnitude of the reference count data. This is done because smaller expression values have higher variance than larger ones. Then we draw fuzzy sets describing the possibility that a certain FC value entails a DE event, given a certain expression value. Combining and ranking the fuzzy descriptions defined in mapping and FC evaluation, we obtain our list of candidate DE events.

Results

We are testing our work on NGS data from Illumina and 454 datasets found on SRS repository. The results are compared to classical FC and p-value estimation of differentially expressed references. As a first result, our model
is able to highlight potential false positives and negatives in DE analysis. As a main advantage emerging from the evaluation of the model, the proposed workflow does not require any normalization step but manages data without strong simplifications or complex probabilistic computations.

Conclusions
The work described in this abstract exploits fuzzy sets to define uncertainty in the evaluation of differential expression on NGS sequencing data. We are studying the first results on control-case datasets. Future developments include the extension of the model in order to cope with more than one sample for condition and with more than two conditions.

nc-aReNA: an integrated web-based platform for non-coding RNA-seq data classification and annotation

Tulipano A(1)(*), De Caro G(1)(*), Consiglio A(1), D’Elia D(1), Grillo G(1), Marinaro M(2), Liuni S(1), Gisel A(1)(*) and Licciulli F(1)(*)
(1)CNR-Istituto di Tecnologie Biomediche, Bari, Italy
(2)ISS-Istituto Superiore di Sanità, Rome, Italy
(*)=Equal contributors
flavio.licciulli@ba.itb.cnr.it

Motivations
RNA-Seq by massively parallel sequencing is a potent way to perform transcriptome and small non-coding RNA (ncRNAs) analyses. Recent reports of the ENCODE project underline that while 80% of the human genome is transcribed, only 2% is protein coding, suggesting that the vast majority of the genome is transcribed as non-protein-coding RNA. We have developed an automated web-based platform, nc-aReNA, for the mapping, classification and annotation of human and mouse ncRNAs from RNA-Seq data. The platform is based on a data-warehouse approach and workflow environment that includes data quality control, genome and nc-RNAome sequence alignment, differential expression profiling analysis and statistics of classified data.

Methods
The nc-aReNA architecture is based on a modular analysis pipeline, with a data-warehouse, for the classification and annotation of small non-coding RNA-Seq data. The pipeline takes in the sequenced reads in FASTQ format. After the initial steps of adaptor removal and quality check, the reads are mapped to an in-house non-redundant ncRNA reference database (http://ncRNAdb.ba.itb.cnr.it) which collects and integrates ncRNA gene lists from MGI (Mouse Genome Informatics) and HGNC (Human Genome Nomenclature Committee), with sequences and biotype annotations from VEGA (Vertebrate Genome Annotation), ENSEMBL, RefSeq, RFam (for tRNA sequence) and miRBase (for miRNA). Currently, the nc-aReNA reference database section contains a total of 62439 (human) and 42330 (mouse) ncRNA sequences classified in 26 biotypes associated to SO terms. As for functional annotation, the data-warehouse contains, among others, information about 410,581 biochemical pathways and 121,579 experimentally validated miRNA target interactions.

Mapped NGS reads are classified in ncRNA categories by using Sequence Ontology (SO). The unmapped reads are aligned to the reference genome and tagged to the corresponding genomic locus. Integrated statistics are used for RPM (Reads Per
address used for the registration in the "monitoring jobs" section. At the end of the analysis the user will receive an e-mail including the URL to download a compressed folder containing the produced results. In the folder it is also included an html report showing general information about the performed analysis and six interactive pie-charts summarizing the inferred taxonomy at different rank levels (from species to phylum). Moreover, a tabular file is supplied in order to allow comparative analysis between different samples by using the BioMaS Post-Processing Tools and METAGENassist.

The Illumina version of BioMaS has been exposed in the Biodiversity Catalogue as a REST service and it is, also, available as Galaxy-based portal at http://galaxy.cloud.ba.infn.it:8080. Moreover, Both the 454 and Illumina BioMaS versions are freely available in a web-service hosted by the ReCaS portal (https://recasgateway.ba.infn.it). A comprehensive documentation about how to register and use both the Galaxy and the ReCas hosted web-services of BioMaS is available at http://testjst.ba.infn.it/openacces/BIOMAS_HELP/BiomasPostProcessingTools.zip.

Conclusions

BioMaS includes a wide range of bioinformatics tools integrated into an automated workflow allowing the user to rapidly obtain a comprehensive and deep view of the microbial taxonomic complexity of the investigated environment. Its user-friendly interface has been designed to enable also users without any computer skills to exploit the BioMaS analysis chain.

Links to workflow and user’s documentation

http://testjst.ba.infn.it/openacces/BIOMAS_HELP/BiomasPostProcessingTools.zip

Differential miRNA-mRNA correlation analysis in colorectal cancer

TM Creanza(1,2), A Piepoli(3), A Andriulli(3), N Ancona(1)

(1) Institute of Intelligent Systems for Automation, National Research Council, CNR-ISSIA, Bari, Italy
(2) Center for Complex Systems in Molecular Biology and Medicine, University of Torino, Torino, Italy,
(3) Department of Medical Sciences, Division and Laboratory of Gastroenterology, IRCCS Casa Sollievo della Sofferenza, San Giovanni Rotondo, Italy.

creanza@ba.issia.cnr.it

Motivations

Colorectal cancer (CRC) is one of the most common cancers worldwide and its molecular biology is one of the most intensively studied. Aberrant microRNA (miRNA) expression is associated with the onset and progression of CRC by regulating the stability and the translation of oncogenes and tumor suppressor genes [1]. Moreover, miRNA-expression profiling has been exploited to predict the functions of the deregulated miRNAs by detecting pathways enriched of their target genes. In cancer, genetic variants in miRNA genes and mRNA targets can affect miRNA-mediated repression [2].

Methods

We studied changes in miRNA-mRNA CRC interactions in terms of differential co-expressions relative to normal condition. The aim was to capture alterations resulting from the aforementioned modifications that influence miRNA control on gene transcription. By using paired miRNA-mRNA expression data, correlations between miRNA and gene expressions were estimated for both tumor and normal tissues. Correlation changes between the two conditions were incorporated into scores of predefined gene sets in order to seek signaling pathways and biological processes with an altered miRNA-mediated control.

Results

Our integrative analysis suggested an alteration in CRC tissues in the interplay between miRNAs and the eukaryotic translation initiation factor 3 (eIF3) which has a central role in recruiting both mRNAs and the cellular translation machinery to form translation initiation complexes [3].
Conclusions

Compared with other miRNA-mRNA integration strategies, our data-driven approach with any a priori information has several advantages: a complete coverage of the human genes on the chip, little bias due to the knowledge obtained from the literature, and the ability to infer condition specific relationships. Unveiling differential miRNA-mRNA co-expression properties allows to gain insights into miRNA-mediated molecular mechanisms underlying the pathogenesis of the disease and may suggest novel drug targets to be validated.

References


Fosso B(1), Santamaria M(1), Marzano M(1), Donvito G(2), Monaco A(2), Notarangelo P(2), Maggi GP(2)(3), Pesole G(1)(4)
(1)Institute of Biomembranes and Bioenergetics, CNR, Bari, Italy.
(2)National Institute of Nuclear Physics, Bari, Italy.
(3)Politecnico di Bari, Bari, Italy
(4)Department of Biosciences, Biotechnology and Biopharmaceutics, University of Bari, Bari, Italy

b.fosso@ibbe.cnr.it

Motivations

Metagenomics, increasingly supported by the progress of High-Throughput Sequencing (HTS) technologies, revolutionized several research fields, such as microbiology, molecular evolution and biodiversity. Currently, a still serious bottleneck is represented by the availability of effective and reliable systems for a friendly and comprehensive large-scale bioinformatic analysis of metagenomic reads. By developing BioMaS (Bioinformatic analysis of Metagenomic AmpliconS) web service we aimed to provide an easy-to-use, versatile and completely automatic workflow for Meta-barcoding data analysis, implementing in a modular framework state of the art tools, from raw HTS data upload and quality assessment to final qualitative and quantitative taxonomic characterization.

Methods

BioMaS has been developed as a modular pipeline based on third-party tools and ad hoc developed Python and Bash scripts. The pipeline application is considerably expensive in terms of computational and storage resources, thus it has been embedded on the Grid/Cloud infrastructure by using the JST (Job Submission Tool) and the web services technologies (REST-SOAP interface).

Results

BioMaS allows the analysis of both bacterial and fungal communities and it is able to process data obtained by Roche 454 or Illumina platforms. The access to the web portals (Galaxy, ReCaS) exploiting web-service tools is allowed through a free registration, by providing the user e-mail address. Before to start the BioMaS computation, the sequencing platform has to be selected, the sequence file submitted and a label associated to the starting job. The process progress can be monitored by simply reporting the same e-mail...
On the Bc2S, user can get access to the resources by means of two web portal: a customized version of Galaxy (Bioinformatics Workflow Management System) and the ReCaS Science Gateway. Within these portals are deployed different user application, for example a modular pipeline, called BioMaS (Bioinformatic analysis of Metagenomic ampliconS). This application provides a comprehensive platform to analyze the huge amount of amplicon-based metagenomic data produced by NGS platforms and can be launch by both web portal.

Conclusions
The versatility of the solutions implemented in BC2S datacenter provides common tools to satisfy the computational requirements of many experiments in the various scientific communities on the Grid/Cloud environment.

In particular, by means tools like JST, it is possible to trade on our infrastructure with workflow manager system o web portal widely used in Life Science experiments.

Global CpG methylation degree brands normal and cancer cells

A De Grassi
Department of Biosciences, Biotechnology and Biopharmaceutics, University of Bari, Italy
anna.degrassi@gmail.com

Motivations
Cytosine methylation at CpG dinucleotides is a well known mechanism of epigenetic regulation of gene expression, which may be altered by a local variation in the methylation pattern of the gene promoter and/or the gene body. Yet, poor is known about the trend of a cell to globally increase or decrease its CpG methylation degree on a genome-wide scale, about the mechanisms that promote this possible trend, and about its consequences on the transcriptome.

Methods
Data for about 50 normal and cancer cell lines were collected from the Encode platform (http://www.genome.gov/encode/), regarding expression array profiles, reduced representation bisulfite sequencing (RRBS), and other genome-wide epigenetic marks. A subset of ~130,000 CpG sites scattered throughout the genome were selected for the high reproducibility among RRBS replicates and were used to estimate the global CpG methylation degree of each cell line. The expression levels of ~20,000 genes from the same cell sets were screened for significant positive or negative correlation with the global CpG methylation degree. Hyper-methylated CpGs in cancer cells were also screened for the presence of other epigenetic marks including the binding to differently modified histones.

Results
Cancer cell lines are distinguishable from primary cell lines with 100% sensitivity and specificity, being the global CpG methylation degree higher in the former versus the latter in all the genome compartments. Several genes were identified whose expression profile positively or negatively correlates with the global CpG methylation increase observed in cancer cells. These genes include known and possibly new markers of cancer cells, encompassing genes with DNA methyltransferase activity, but not genes with DNA demethylating activity. Further, those CpG sites that are hyper-methylated in
cancer cells and hypo-methylated in normal cells are both marked by the same histone modification in both the cell line types.

Conclusions

The present results underpin three major conclusions. First, cancer cells are marked by a basal and homogeneous CpG hyper-methylation status. Second, this status seems to be promoted by the over-expression of DNA methyltransferases, and not by the transcriptional repression of DNA demethylator genes. Third, a specific histone mark occurs at CpG sites that are methylated only in cancer cells, suggesting that another epigenetic mark may have predated and maybe favored the CpG methylating process itself.
Extra filtering and other options are also feasible, such as duplicate reads removal and addition of heteroplasmic variants in the consensus genome above a threshold chosen by the user.

Results

The tool provides a summary file showing statistics about the coverage of reconstructed mitochondrial genomes (for NGS data), haplogroup(s) prediction and annotation of prioritized variants. A compressed folder containing all output files generated by the pipeline can be downloaded once the analysis is completed, including a VCF file, with the corresponding heteroplasmic fraction and confidence interval as genotype-specific meta-information.

Conclusions

MToolBox as a command line tool was successfully used to annotate all mitochondrial variants identified in the 1000 Genomes dataset, showing heteroplasmy spread in healthy samples [2].

The full Graphical User Interface (GUI) will allow also clinicians, and not only computer specialists, to extensively use the tool for mitochondrial genome analyses.

References


ComiRNet: a new database using an integrative data-mining approach for the discovery of miRNA-gene regulatory networks

G Pio(1), M Ceci(1), D Malerba(1), D D’Elia(2)
(1)Department of Computer Science, University of Bari Aldo Moro, Bari, Italy
(2)CNR, Institute for Biomedical Technologies, Bari, Italy
domenica.delia@ba.itb.cnr.it

Motivations

Computational approaches have been proven to be fundamental in the miRNA research for both gene-specific and large-scale predictions of miRNA targets, for the formulation of new functional hypothesis on their biological role, for gene network discovery and to guide experimental validations. However, their effectiveness is negatively affected by high uncertainty of micro RNA (miRNA) gene target predictions and by the complexity of rules governing miRNA functional targeting, whose mechanisms still remain elusive. In order to improve predictions of miRNA targets and to support the elucidation of miRNA functional role in the context of gene regulatory networks, we have developed a new two-stepped computational approach based on: i) a semi-supervised ensemble-based classifier for the prediction of miRNA target interactions (MTIs) [1] and, ii) a biclustering algorithm (HOCCLUS2) for the prediction of miRNA-gene regulatory networks (MGRNs) [2]. Data produced are available at ComiRNet (Co-clustered miRNA Regulatory Networks: http://comirnet.di.uniba.it:8080/), a user-friendly web-based system providing efficient query, retrieval, export, visualization and analysis of predicted MTIs and MGRNs.

Methods

In the first step, a semi-supervised ensemble-based classifier is learned from both experimentally validated interactions (positively labelled examples), extracted from miRTarBase [3], and miRNA gene target predictions (MTIs), returned from several prediction algorithms (unlabelled examples) and extracted from mirDIP [4]. This classifier acts as a meta-classifier of unlabelled examples. As a result of the first step, a unique (meta-)prediction score is available for all possible interactions. In the second step, these prediction scores are used to identify miRNA-gene regulatory networks (MGRNs) through the biclustering algorithm HOCCLUS2. HOCCLUS2 exploits the large set of produced predictions, with the associated probability, to extract a set of overlapping and hierarchically organized biclusters each one representing putative MGRNs. The construction of the hierarchy is performed by an iterative merging, considering both distance and...
density-based criteria. Extracted biclusters are also ranked on the basis of the p-values obtained by the Student’s T-Test which compares intra- and inter-functional similarity of miRNA targets, computed on the basis of the gene classification provided in Gene Ontology (GO) [5]. The ComiRNet database relies on PostgreSQL DBMS, while the web-based platform is built through the Play 2.2 Java framework and the Cytoscape library.

Results
ComiRNet currently stores approximately 5 million predicted interactions between 934 human miRNAs and 30,875 mRNAs and 15 different bicluster hierarchies. The ComiRNet web interface allows users to perform extraction and visualization of single interactions (with the score/probability assigned by the learning algorithm) and of biclusters of interest, as well as to easily browse whole biclusters hierarchies. The interface for the analysis of biclusters also provides a graph-based visualization of the predicted miRNA-gene interaction networks. Biclusters hierarchy browsing (i.e., navigation among parents and children biclusters) helps to identify intrinsic and functional relationships between different miRNAs and their predicted functional co-targeting on different groups of genes.

Conclusions
Based on the principles of our approach, genes in a bicluster are likely to function together as a network and miRNAs in the same bicluster are likely to cooperatively target groups of networked genes. The use of computational predictions in place of only experimentally validated interactions offers the possibility to detect single interactions and regulatory modules that would be otherwise impossible to reconstruct by considering only experimentally validated interactions, which are strictly dependent on the cell type and experimental conditions used. This paves the way to the systematic use of ComiRNet for i) a comprehensive analysis of cooperative targeting of miRNAs of interest; ii) the discovery of unknown miRNA and gene functions, on the basis of the functional similarity suggested by ComiRNet biclustering; iii) the discovery of unknown miRNA targets which could be worth to be experimentally validated. This possibility is due to the ComiRNet ability to associate objects that are apparently not related.

MToolBox, an MSeqDR website tool for human mitochondrial DNA analysis from high throughput and Sanger sequencing data

M.A. Diroma (1), L. Shen (2), C. Calabrese (3), D. Simone (1), M. Santorsola (4), X. Gai (2), M. Attimonelli (1)
(1) Department of Biosciences, Biotechnologies and Biopharmaceutics, University of Bari, Italy
(2) Department of Ophthalmology, Massachusetts Eye and Ear Infirmary, Harvard Medical School, Boston, USA
(3) Department of Medical and Surgical Sciences, University of Bologna, Italy
(4) Department of Sciences and Technologies, University of Sannio, Benevento, Italy
mariangeladiroma@gmail.com

Motivations
The Mitochondrial Disease Sequence Data Resource (MSeqDR) Consortium (https://mseqdr.org/) was established in 2012 as a central data resource for human mitochondrial DNA (mtDNA) analyses for mitochondrial disease community. MSeqDR provides access to a wide range of web-based bioinformatics tools, including MToolBox, a pipeline formerly developed as a stand-alone tool [1] for mtDNA assembling and analysis from high throughput and automated sequencing data. The software was developed to accurately measure heteroplasmy and easily recognize the most functionally important mitochondrial variants.

Methods
A user-friendly graphical interface was developed (https://mseqdr.org/mtoolbox.php) to easily manage the analysis. A user can upload data files in FASTQ, SAM/BAM or FASTA formats. The default filtering based on quality score, coverage and exclusion of reads mapping on both nuclear and mitochondrial sequences is carried out to obtain variant calling and reconstruction of consensus genome. The tool uses hg19 as nuclear and rCRS or RSRS as mitochondrial reference sequences, at user’s choice.

After haplogroup assignment for each analyzed sample, annotation and prioritization of detected variants are performed through variants recognition by both the mitochondrial reference sequences and a related macro-haplogroup consensus sequence.