Inteligencia Artificial aplicada a BioData

Artificial Intelligence applied to BioData

Volumen 1 (2021) / *Volume 1 (2021)*

18-19 November, 2021

BOOK OF ABSTRACTS

Actas / *Proceedings*
del Taller / *Workshop* de la
Red Iberoamericana *RiaBio*

RiaBio

Actas del Taller de la Red Iberoamericana RIABIO sobre Inteligencia Artificial aplicada a BioData / Editores: Flavio E. Spetale; Debora Arce; Elizabeth Tapia; Javier De Las Rivas. 1ra Edición. RIABIO Editora. Salamanca, 2021.

Libro digital, PDF . Archivo Digital: descarga y online

ISBN XXX

1. RIABIO Editora. Spetale F. E.; Arce D.; Tapia E.; De Las Rivas J.
Título: **Actas del Taller de la Red Iberoamericana RIABIO sobre Inteligencia Artificial aplicada a BioData / Proceedings of the Workshop of RIABIO IberoAmerican Network on Artificial Intelligence applied to BioData**

Diseño interior y de tapa RIABIO Editora

ISBN XXX

RIABIO Editora
Campus Miguel de Unamuno s/n – 37007 Salamanca, España
coreRiaBio@cifasis-conicet.gov.ar
Impreso en España, México y Argentina.

Red Iberoamericana de Inteligencia Artificial para Big BIOdata (RiaBio)

Dra. Elizabeth Tapia
Coordinadora/directora

Dr. Javier De Las Rivas
Co-coordinador/co-director

Dra. Maribel Hernández Rosales
Co-coordinadora/co-directora

Comité Científico/Organizador

Dra. Elizabeth Tapia
Dr. Flavio E. Spetale
Dr. Javier De Las Rivas
Dra. Maribel Hernández Rosales
Dr. Jorge Valdés

Dra. Alejandra Medina
Dra. Débora Arce
Dr. Cesar Bonavides
Dr. Luis Tataje
Dra. Fiorella Cravero
BSc. Cleidy Osorio Mogollón

Coordinación de edición

Dr. Flavio E. Spetale
Dr. Javier De Las Rivas
Dra. Elizabeth Tapia

Auspicio

CYTED
Centro de Investigación de Cáncer (CiC-IMBCC, CSIC/USAL)

SUMARIO

Prólogo

El presente ejemplar marca la concreción de la primera edición de nuestro Taller de la Red Iberoamericana RiaBio. Docentes, investigadores y empresas han acompañado y apoyado este esfuerzo conjunto como lo atestiguan los 20 resúmenes que componen el presente volumen.

RiaBio tiene como objetivo central la creación de una red de grupos de investigación y entidades iberoamericanas (públicas y privadas) que permita afrontar los desafíos y oportunidades emergentes de la creciente fusión de innovaciones en los campos de la Inteligencia Artificial (IA) y la ciencia de datos en Biología (DataScience).

Programa / Program

Día 1º / Day 1: 18. Nov.2021 Jueves / Thursday			
Hora / Time from 16:00 to 20:00 (CET, Madrid) = 12:00-16:00 (- 4HRS Buenos Aires/Santiago/Montevideo) = 10:00-14:00 (- 6HRS Bogotá) = 09:00-13:00 (- 7HRS Ciudad de México)			
16:00 - 16:15	Presentación del Primer TALLER RIABIO: **Elizabeth Tapia** y **Javier De Las Rivas** tapia@cifasis-conicet.gov.ar y jrivas@usal.es		
16:15 - 16:45	Keynote Speaker 1	**Roberto López (ES)** robertolopez@artelnics.com	Aplicaciones de Inteligencia Artificial en Medicina
16:45 - 17:00	1ª Comunicación Oral	**Óscar González-Velasco (ES)** oscargv@usal.es	Deep Learning in clinical diagnosis: primary site tumor identification of metastatic cancers and explainable machine learning as a tool for biomolecular understanding
17:00 - 17:15	2ª Comunicación Oral	**Mauricio Vueltiflor (MX)** mauriciovueltiflor@gmail.com	Modelos computacionales de clasificación aplicados al estudio de Transferencia Horizontal de material genético
17:15 - 17:30	3ª Comunicación Oral	**Amilcar Meneses Viveros (MX)** amilcar.meneses@cinvestav.mx	Heterogeneous computing applied to bioinformatics problems
17:30 - 17:45	4ª Comunicación Oral	**José Manuel Gilpérez Aguilar (ES)** josemanuel.gilperez@uclm.es	Machine Learning model for epigenetic pattern recognition and prediction of enhancers and promoters
17:45 - 18:00	5ª Comunicación Oral	**Leandro Murgas Saavedra (CL)** leandro.murgas@mayor.cl	Machine Learning approach to determine relationships between epigenetics marks
18:00 - 18:15	Café Descanso / Coffe Break		
18:15 - 18:30	6ª Comunicación Oral	**Alberto Berral-Gonzalez (ES)** aberralgonzalez@usal.es	Bioinformatic algorithm for the prediction of patient risk based on omic data and survival
18:30 - 18:45	7ª Comunicación Oral	**Fernando Fontove** fernando.fontove@c3consensus.com	Gene Finder: a tool for small gene discovery
18:45 - 19:00	8ª Comunicación Oral	**Alejandra Serrano-Rubio (MX)** angelica.serrano@cinvestav.mx	Gene expression analysis through parallel Nonnegative Matrix Factorization
19:00 - 19:15	9ª Comunicación Oral	**Flavio E. Spetale (AR)** spetale@cifasis-conicet.gov.ar	GO Deep: AI Annotation of lncRNAs
19:15 - 19:30	10ª Comunicación Oral	**Octavio Zambada Moreno (MX)** octavio.zambadam@cinvestav.mx	Multi-omics analysis to determine the master transcription factors regulators in tomato response to PSTVd infection
19:30 - 19:45	11ª Comunicación Oral	**Martina Fernández (AR)** martinafernandez1997@gmail.com	Development of a highly multiplexed full-length 16S rRNA sequencing protocol for the gut microbiota of low-temperature tolerant Pacú
19:45 - 20:00	12ª Comunicación Oral	**Pedro Sepúlveda (CL)** pyter.sr@gmail.com	Genotypic and transcriptional regulatory diversity in the extremely acidophilic Acidithiobacillus genus: insights into anaerobic metabolism
20:00 - 20:20	Keynote Speaker 2	**Daniel Aguayo (CL)** daniel.aguayo@unab.cl	Cosechando rendimiento con Inteligencia Artificial

Día 2º / Day 2: 19. Nov.2021 Viernes / Friday

Hora / Time from 16:00 to 20:00 (CET, Madrid) = 12:00-16:00 (- 4HRS Buenos Aires/Santiago/Montevideo) = 10:00-14:00 (- 6HRS Bogotá) = 09:00-13:00 (- 7HRS Ciudad de México)

16:00 - 16:15	Presentación de Actividades y Objetivos de la Red RIABIO – CYTED: **Jorge Valdés** y **Javier De Las Rivas** jorge.valdes@gmail.com y jrivas@usal.es		
16:15 - 16:45	Keynote Speaker 3	**Elizabeth Tapia (AR)** tapia@cifasis-conicet.gov.ar	Bayesian probabilistic inference and the development of a highly multiplex long read sequencing protocol for SARS-CoV-2 genomes
16:45 - 17:00	1ª Comunicación Oral	**Mónica Padilla-Gálvez (MX)** mpadilla@lcgej.unam.mx	Uncovering the regulatory network that modules the transcriptional response to SARS-CoV-2 infection in humans
17:00 - 17:15	2ª Comunicación Oral	**Sebastián Contreras-Riquelme (CL)** contrerasriquelme.sebastian@gmail.com	Deciphering the relationship between epigenetic marks and Transcription Factor Binding
17:15 - 17:30	3ª Comunicación Oral	**Camilo Villaman (CL)** c.villaman@gmail.com	CTCF binding prediction using genomic and epigenomic features
17:30 - 17:45	4ª Comunicación Oral	**Néstor Diaz (CO)** nediaz@unicauca.edu.co	Imbalanced fetal state classification from Cardiotocograms
17:45 - 18:00	5ª Comunicación Oral	**Enrique De La Rosa Morón (ES)** enriquedlrm98@gmail.com	Using current bioinformatics tools to analyze single-cell transcriptomic data
18:00 - 18:15	Café Descanso / Coffe Break		
18:15 - 18:30	6ª Comunicación Oral	**Natalia Alonso Moreda (ES)** id00740098@usal.es	Mixture deconvolution methods of transcriptomic data to identify cellular composition and markers
18:30 - 18:45	7ª Comunicación Oral	**Margot Paulino Zunini (UY)** margot@fq.edu.uy	Articulación de estrategias “Ligand Based”, “Structure Based Drug Design” y de Inteligencia Artificial aplicadas al diseño innovador de compuestos bioactivos
18:45 - 19:00	8ª Comunicación Oral	**Pablo Garcia Briosso (UY)** pcgarcia@windowslive.com	Representación numérica de ligandos y sitios de anclaje para el diseño de modelos basados en inteligencia artificial
19:00 - 19:15	9ª Comunicación Oral	**Maribel Hernández Rosales (MX)** maribel.hr@cinvestav.mx	Differentially expressed intronless genes across cancer highlight their functional role in epigenetics
19:15 - 19:30	10ª Comunicación Oral	**Katia Aviña Padilla (MX)** katia.avinap@cinvestav.mx	Eukaryotic intronless genes: insights into their functional significance and evolution
19:30 - 19:45	11ª Comunicación Oral	**Verónica Latapiat (CL)** veronica.latapiat@mayor.cl	Study of stratification based on transcriptomic in subjects with Alzheimer’s disease using individualized co-expression networks.
19:45 - 20:00	12ª Comunicación Oral	**Sandra Arancibia Opazo (CL)** sandraaran@gmail.com	Analysis of transcriptional changes associated with CRE elements in a murine model of Huntington's disease due to interaction of CBP with mutant Huntingtin protein
20:00 - 20:30	Foro RIABIO - Discusión General / Cierre del Evento		

Applications of Artificial Intelligence in Medicine

Roberto Lopez[1]

1. Neural Designer

Background:
Artificial Intelligence is transforming the practice of medicine. For example, it allows diagnosing patients' diseases earlier, predicting their evolution, and providing them with personalized treatments.

Results:
In this work, we present the three most common types of artificial intelligence applications in medicine: medical diagnosis, medical prognosis and medical treatment. We also work out an example for each, using the data science and machine learning platform Neural Designer. As an example of medical diagnosis, we train a neural network to specify dermatological diseases from clinical and histopathological variables. To illustrate medical prognosis, we build a predictive model to assess whether a patient will develop diabetic retinopathy or not, based on age, diastolic and systolic pressure and cholesterol level. Finally, we include an example of medical treatment that tries to determine the best medicine for a patient who has undergone colon cancer surgery.

Conclusions:
This work shows that artificial intelligence has very diverse applications in medicine. However, we solve all of them in a very similar way. In addition, thanks to dedicated tools, you don't need to be a math or programming expert to build data-driven models.

Deep Learning in clinical diagnosis: primary site tumor identification of metastatic cancers and explainable machine learning as a tool for biomolecular understanding

Oscar González-Velasco[1], Javier De Las Rivas[1]

1. Cancer Research Center (CiC-IBMCC, CSIC/USAL/IBSAL), Consejo Superior de Investigaciones Científicas (CSIC), University of Salamanca (USAL) & Instituto de Investigación Biomédica de Salamanca (IBSAL), 37007 Salamanca, Spain

Background:

Cancers of Unknown Primary (CUP) represent a heterogeneous group of metastatic cancer that are poorly differentiated and whose primary site is not known at the time of diagnosis. CUPs are extremely complex and poses a difficult challenge for both biomolecular understanding of causality and, consequently, effective diagnosis and treatment. Advances in recent years and access to larges collections of datasets is changing this situation rapidly: the number of research projects on cancer has increased exponentially, and those papers that involved machine learning and artificial intelligence techniques applied to large collection of cancer samples, has grown especially fast.

Results:

Using a large cohort of cancer RNA-Seq samples and healthy tissue biopsies from 27 different primary sites, we have built a Convolutional Neural Network (CNN) model that processes a matrix composed by the mRNA expression of a set of human genes, as a result the model gives the predicted probability for the sample of being of any of the possible primary sites. An additional model using a data subset of samples was built for predicting progression and recurrence after treatment. The input expression matrix is composed of 4489 genes and has been build using a set of transcription factors and its target genes, the construction of the matrix has been arranged following a specific order maximizing the number of shared targets using network graphs and the Kamada-Kawai algorithm for visually representing graphs, thus optimizing the data for the use of convolutional neural networks. Additionally, we have use explainable machine learning methods to weight the contribution of the targets to the given prediction, allowing us to select biomarkers for study as the difference between metastatic and primary tumor as well as the cause of progression and recurrence. Our CNN model shows a 97% accuracy using the validation RNA-Seq samples, and more importantly, it shows and accuracy ranging from 92% to 97% on new external datasets, including distant metastatic samples from kidney and lung.

Conclusions:

These highly accurate predictions show the great potential of the use of AI in real clinical scenarios, and the use of Deep Learning as a tool for novel analysis of markers, and the interaction between transcription factors and their targets, as a mean to study biomolecular changes on diseases.

Modelos computacionales de clasificación aplicados al estudio de transferencia horizontal de material genético

Mauricio Vueltiflor[1]

1. Cinvestav Unidad Irapuato, Mexico

Abstract:

La transferencia horizontal de material genético es un evento de suma importancia para la historia evolutiva de las especies. Sin embargo, los métodos bioinformáticos para detectarla se basan mayormente en características funcionales del ADN codificante y por ello se enfocan en la transferencia horizontal de genes, pero hay evidencia de transferencia horizontal de ADN no codificante. Además, se sigue poniendo en evidencia la importancia del ADN no codificante con cada nueva funcionalidad biológica que se descubre en él. En este trabajo se propone el uso de características del ADN que reflejan magnitudes termodinámicas, fisicoquímicas y estructurales las cuales dependen únicamente de la secuencia primaria de ADN y no de alguna característica relacionada con la codificación para proteínas. Con estas magnitudes empleamos métodos computacionales para clasificar material genético entre categorías taxonómicas y, específicamente, entre especies. Con este mecanismo de clasificación de material genético se propone usarlo para detectar no solo eventos de trasferencia horizontal de genes, sino también de transferencia horizontal de material genético en general.

Heterogeneous computing applied to bioinformatics problems

Amilcar Meneses Viveros[1]

1. Departamento de Computación, Cinvestav-IPN, México

Background:
Several problems in bioinformatics require great computing power, either due to the large amount of data that must be processed or due to the complexity of the algorithms used. Computers with large storage capacities, memory and various processing units are used to process bioinformatics data. The processing units can be homogeneous or heterogeneous, that is, the processing units have the same processing capacity or not. Using an AI approach, it is possible to have an automatic mechanism that selects the best processing units to perform some computing tasks.

Results:
Using Machine Learning techniques, we have managed to accelerate some parallel problems in heterogeneous architectures. Small heterogeneous parallel libraries have been developed and Machine Learning techniques have been incorporated for scheduling tasks between the various processing units.

Conclusions:
The size of the problem is the most relevant attribute for machine learning strategies to obtain better accelerations.

Machine learning model for epigenetic pattern recognition and prediction of enhancers and promoters

Daniel Rodríguez[1], Enrique Pérez[1], Juan Martínez[2], Lucía Ramírez-Navarro[2], José Manuel Gilpérez[1], Alejandra Medina-Rivera[2]

1. EIIA de Toledo, Universidad de Castilla La Mancha, España
2. LIIGH-UNAM, México

Background:
The characterization of enhancers and promoters is of great interest to determine the activity of these non-coding elements in different cell lines. Currently, the existence of massively parallel tests (i.e., STARR-seq) for the characterization of enhancers enables us to have a huge volume of data that allows us to create machine learning models for the prediction of these elements. However, to date, these models are difficult to reproduce and lack adequate optimization that hampers usability. Moreover, new methods (i.e., CapSTARR-seq) have helped identify promoters with enhancer function, known as ePromoters. Likewise, the same model allows the integration of additional epigenetic signals, as well as the determination of the double function of promoter and enhancer of some of these elements.

Results:
The main objective is the application of a machine learning model to identify regions of the genome with regulatory activity such as enhancers, promoters or dual functions. We start from a previous model (Gerstein Lab) on which different modifications are made to optimize its operation. In a first phase, the appropriate signals for training or metaprofiles, are generated by processing signals from different histones such as H3K27ac or DNase hypersensitive sites (DHSs) that overlap STARR-seq regions. Then, using these signals, a predictive model based on Support Vector Machine (SVM) is applied first. Using the same signals, it is intended to develop a new model based on deep convolutional neural networks that improves predictive capabilities while aiming at improving prediction of ePromoter regions. In parallel, we study whether conservation can be introduced as a new characteristic associated with the epigenetic function. For this, various conservation measures are used. In this way, we verify that the regions with the highest genomic conservation coincide to a certain extent with those that present promoter, enhancer or ePromoter behavior. Then, this new feature is included in the previous machine learning model.

Conclusions:
We have developed a machine learning model applicable to the study of epigenetic function, based on a previous model, which we have optimized by improving its functionality. Likewise, we have added conservation as a useful characteristic to identify this epigenetic function. Soon, we will add improvements to the model by applying convolutional neural networks.

Machine Learning approach to determine relationships between epigenetics marks

Leandro Murgas Saavedra[1,2], Mauricio Sáez[3], Alberto J. Martin[2]

1. Programa de Doctorado en Genómica Integrativa, Vicerrectoría de Investigación, Universidad Mayor
2. Network Biology Lab., Centro de Genómica y Bioinformática, Facultad de Ciencias, Universidad Mayor
3. Chromatin, Epigenetic, and Neuroscience Laboratory, Centro de Genómica y Bioinformática, Facultad de Ciencias, Universidad Mayor

Background:

Different complex genetic diseases are related to variations in epigenetics that lead to a change in chromatin states, being reflected in highly de-regulated transcription. In some types of cancer, such as colorectal cancer, altered patterns of histone modifications and DNA methylation have been identified and are deemed to be one of the causes behind the altered transcriptional landscape of tumors. Importantly, transcriptional activity of chromatin can be classified into different states depending on patterns of epigenetic marks, but to carry out accurate state assignment many different marks are needed. For this reason, knowing what state each chromatin region is in and observing the changes in these states is a promising source of information to better understand complex pathologies. Thus, the purpose of this work is to reduce the number of epigenetic marks by means of machine learning tools to find non-obvious redundancy relationships between different epigenetic marks, allowing to reduce the number of experiments required to determine chromatin states, increasing the feasibility of this type of analysis.

Results:

In this work, various models of the Random Forest type were trained for different epigenetic marks, using data from ChIP-seq experiments of histone modifications, varying parameters of the algorithm and the way in which the training data were provided, to find the optimal way to represent the information. In this way, it was possible to determine various relationships between epigenetic marks, being able to identify which of them have a greater relevance to be able to predict others.

Conclusions:

The results obtained to date since this work is currently in progress. Our findings indicate that the prediction of several types of epigenetic marks using information from ChIP-seq experiments of other histone modifications, is possible. Our preliminary results have also proven accurate enough to allow robust chromatin state assignment by combining predicted marks with those from experimental results, increasing in this way the number of experiments required to study alterations on chromatin states without performing many costly experiments.

Bioinformatic algorithm for the prediction of patient risk based on omic data and survival

A. Berral-Gonzalez[1], S. Bueno-Fortes[1], J.M. Sanchez-Santos[1], M. Martin-Merino[1] and J. De Las Rivas[1]

1. Bioinformatics and Functional Genomics Group - Cancer Research Center (CJC-JBMCC, USAL/CSJC/JBSAL), Salamanca, Spain

Background:

Nowadays, with the support of different omic technologies, data-driven medicine makes possible to study diseases with different approaches, allowing to assess the role of genes as possible molecular biomarkers of risk, prognosis or patient outcome. Therefore, to achieve a better prognosis or prediction of the disease, the discovery and validation of survival biomarkers associated with a specific phenotype or clinical variables is a critical step. In addition, this will allow the development of personalized treatments and precision medicine, based on patients accurate risk prediction and stratification. Currently, discovering gene markers for survival or evaluating the prognostic capacity of specific genetic signatures is quite a complex job. Furthermore, it is common that certain markers found are not reproducible nor robust, and the gene signatures discovered cannot be correlated well with the clinical phenotypes or with the stages and outcome of the disease. Besides, there are not many integrated tools that provide molecular-based patient risk and survival assessment.

Results:

The developed algorithm provides an integrated set of functions to analyze survival and provide patient risk predictions based on genetic signatures. The tool allows: (i) Discovery of lists of genes associated with survival in diseases based on omic data (expression or activity), in a robust and reproducible way (geneSurv); (ii) Discovery of genetic markers by identifying the association of gene expression (or similar gene signal) with clinical variables or phenotypic characteristics (genePheno); (iii) Construction of robust patient risk predictors based on gene signatures using univariate and multivariate approaches (patientRisk).

Conclusions:

Applying the algorithm developed to real biomedical data in a reproducible and robustly way, we have managed to obtain gene signatures linked to survival. Furthermore, these signatures allow patients to be stratified into high-risk and low-risk.

Gene Finder: a tool for small gene discovery

Fontove Fernando[1], Angulo Carlos[1], Pichardo Israel[2]

1. C3 Consensus
2. Harvard Medical School

Background:
The human genome has been thoroughly analyzed to discover genes or Open Reading Frames (ORFs), traditionally the tools used discarded small sequences as false positives. In the past years, several small ORFs (smORFs) have been discovered and have exhibited properties of great interest, such as exploiting blind spots in the immune system which is of great impact for medicine. Of the tools used for gene discovery, PhyloCSF stands out in the field of comparative genomics. It trains two models for homologous sequence evolution: for gene encoding sequences and for non-coding sequences. The training data comes from homologous regions of diverse species. These models are then used to calculate the likelihood a given set of sequences evolved under a gene-coding or non-coding model and makes a prediction. The downside of this algorithm is the large runtime that prohibits large scale analysis.

Results:
In this work we developed a platform that is capable of genome wide scans to look for smORFs. Our tool's main engine is a reimplementation of PhyloCSF, together with additional tools for sequence extraction, cross referencing with diverse databases: sequence conservation across species, known genes, RNA seq and proteomics. We use the trained models provided by PhyloCSF and make an ad hoc implementation of the evaluation function for three species.
Additionally, we provide a graphical user interface for analysis and reporting that allows for non-expert users.

Conclusions:
Using this platform, we were able to do a full scan of the human genome in under a week's processing time. This analysis yielded 4,716 smORF encoding peptides.

Gene expression analysis through parallel Nonnegative Matrix Factorization

Alejandra Serrano-Rubio[1], Guillermo B. Morales-Luna[1], Amilcar Meneses-Viveros[1]

1. Computer Science Department, CINVESTAV-IPN, México City, México

Background:

Genetic expression analysis is a principal tool to explain the behavior of genes in an organism when exposed to different experimental conditions. In the state-of-art, many clustering algorithms have been proposed. It is overwhelming the amount of biological data whose high-dimensional structure exceeds mostly current computational architectures. Computational time and memory consumption optimization actually become decisive factors in choosing clustering algorithms. We propose a clustering algorithm based on Nonnegative Matrix Factorization and K-means to reduce data dimensionality but preserving the biological context and prioritizing gene selection, and it is implemented within parallel GPU-based environments through the CUDA library. A well-known data set is used in our tests and the quality of the results was measured through the Rand and Accuracy Index. The results showed an increase in acceleration of 6.22x compared to the sequential version. The algorithm is competitive in biological data sets analysis and it is invariant with respect to the classes number and the size of the gene expression matrix.

GO Deep: AI Annotation of lncRNAs

García-Labari I.[1], Spetale F.E.[1, 2], Iglesias N.[1, 2], Murillo J.[1,2], Angelone L.[1, 2], Bulacio P.[1,2], Tapia E.[1, 2]

1. CIFASIS-CONICET-UNR
2. Facultad de Ciencias Exactas, Ingeniería y Agrimensura, UNR

Background:

Gene Ontology (GO) provides access to computable knowledge about genes and gene products. Most of the GO annotation tools developed during the past twenty years focus on protein coding genes known to encode their functionality on their primary sequence. More recently, the fundamental role of non-coding genes in the regulation of protein coding genes has been definitely established. Among non-coding genes, those encoding lncRNA products (> 200 nt) are particularly suitable for their in-silico annotation by Machine Learning methods. In a recent contribution, we showed that access to lncRNA secondary structure information enables their automatic GO annotation. We note, however, that to go deeper in the annotation of lncRNAs, an improved characterization - beyond that provided by naive Kmers - of their secondary structure information is required. Here, we present preliminary results on a novel GO annotation method for lncRNAs where deep learning overcomes the need for expert characterization of primary and secondary structure information.

Results:

We built upon the pipeline described in Spetale et al. 2021 where a hierarchical distributed approach for the supervised GO annotation of lncRNAs was presented. Briefly, the GO graph induces a set of binary SVM predictors of individual GO terms. These predictors provide raw, likely inconsistent, GO annotations for query lncRNA sequences. An instance of the belief propagation algorithm for graphs with cycles, a workaround solution for distributed reasoning in artificial intelligence, leverages raw GO annotations taking into account GO relationships among GO terms and the confidence of raw GO annotations. Individual SVM predictors are trained with lncRNA data from a curated repository. The training process requires the expert, non trivial, characterization of primary sequence data and associated models for candidate secondary structures. Here, we overcome such a need for expert characterization by means of deep learning. At each level of the GO graph, a multiclass CNN is introduced to provide raw GO annotations for the corresponding GO terms. From top to bottom of the GO, the process is repeated until scarcely populated GO terms emerge (less than 500 annotated lncRNAs). At this point, individual GO term predictions are obtained from SVM predictors. Experimental results on GO subsets for zebrafish and human lncRNAs confirm the power of deep learning to accurately predict general GO terms at the top GO levels, a feature that boosts annotations at deeper GO levels.

Conclusions:

The introduction of deep learning processing at the top GO levels removes spurious GO lncRNAs annotations introduced by SVM counterparts. As a result, we can go deeper in the annotation of lncRNAs avoiding the need to analyze confusing GO annotation branches at the posterior expert visualization analysis.

Multi-omics analysis to determine the master transcription factor regulators in tomato response to PSTVd infection

Aviña-Padilla Katia[1], Herrera-Oropeza Emilio[2], Zambada-Moreno Octavio¹, Hernández-Rosales Maribel[1]

1. Cinvestav-Irapuato, México
2. King's College London
3. Instituto de Neurobiología UNAM

Background:

Viroids are minimal pathogens consisting of non-coding RNA that cause multiple diseases in agronomic interest crops. Symptoms associated with viroid infection are linked to developmental alterations due to genetic regulation. The transition from vegetative growth to reproductive development requires gene network coordination, where transcription factors (TFs) act as essential organ morphogenesis components. In tomato host, the infection symptoms by Pospiviroid species include dwarfism, reduction in vigor, abortion of flowers, and reduced size and number in fruits. In order to find key TF, known as master regulators, of the transition of a healty plant to an affected by PSTVd one, we adopted an omics approach centered on transcriptomics data (RNAseq) and gene regulatory networks deconvolution, with the propose of characterize these and identify in which biological processes they are implied that may help the developing of viroid infection symptoms.

Results:

We obtained a gene regulatory network in which all the TFs known in tomato to the date were identified along with their target genes (regulon), in addition using the master regulator analysis algorithm we elucidated which of these regulons were the most differently expressed, thus, the most probable of being responsible of the PSTVd infection symptoms.

Conclusions:

Overall, our results revealed three particular TF families, bHLH, MYB and ERF, which regulate genes among conditions that are involved in molecular mechanisms underlying distinct biological processes. To the best of our knowledge, this work represents the first approach to study the regulatory role of master regulators at the transcriptional and post-translational level in the molecular mechanism of viroid pathogenesis.

Development of a highly multiplexed full-length 16S rRNA sequencing protocol for the gut microbiota of low-temperature tolerant Pacú

Martina Fernandez[1], Ignacio García Labari[2], Victoria Posner[3], Sofia Lavista Llanos[2], Felipe del Pazo[3,4], Alan Marín[3], Florencia Mascali[3,4], Andrés A. Sciara[3,4], Gabriela V. Villanova[3,4], Juan Rubiolo[3,4]

1. Facultad de Ciencias Bioquímicas y Farmacéuticas, UNR
2. Argentag
3. Laboratorio Mixto de Biotecnología Acuática, Centro Científico Tecnológico y Educativo Acuario del Río Paraná
4. CONICET

Background:

The composition and function of animal gut microbiomes, key regulators of host physiology, is affected by temperature variation. Little is known about this in poikilothermic vertebrates, whose physiology is also strongly influenced by environmental temperature. Here, we characterize the gut microbiome of pacú Piaractus mesopotamicus, an ectothermic Neotropical fish species ranking first in farming production in Argentina. Taking into account that there is increased evidence that gut microbiome disturbance may be the indirect way by which temperature impacts the animal fitness, we aim to identify the microbial species present in biological samples of Pacú fish sensitive and tolerant to low temperatures. Toward this goal, we designed a highly multiplexed protocol 16S rRNA microbiome profiling of Pacú gut based on long read sequencing (LRS) technologies.

Results:

Aiming the application of the sequencing protocol in Pacú farming production, we rely on the MinION™ LRS platform (Oxford Nanopore Technologies). To overcome the limits of existing barcoding methods for LRS, both in the number of unique barcodes and in the misassignment of samples due to the challenging indel errors, we rely on a novel family of barcodes, called NS-watermark comprising 4096 members of 36 nt each, described previously by Ezpeleta J. et al 2017. Hence, we developed a two-step PCR amplification and barcoding protocol for the full-length 16S rRNA (~1,500 bp). To evaluate its taxonomic resolution, mock community samples as positive controls were used. Negative controls evidenced the presence of exogenous DNA contamination. To screen this background DNA contamination, we performed nine mock microbial dilution series that were sequenced together with their negative control (blank sample). The dual symmetric NS-watermark barcoding scheme implemented by our protocol ensured a fine control of contamination due to crosstalk multiplexing artifacts. This enabled the confident analysis of demultiplexed reads for meaningful downstream analysis, including the identification and statistical removal of contaminant DNA taxa whose frequencies inversely correlated with sample DNA concentration in mock dilution series and were simultaneously present in negative controls.

Conclusions:

Control of background DNA contamination is a critical issue in 16S rRNA microbiome profiling whose importance is just being recognized. Here, we presented preliminary results on contamination control in highly multiplexed 16S rRNA microbiome studies performed in the MinION LRS platform for which the straightforward elimination of contaminating reads should be carefully managed.

Genotypic and Transcriptional Regulatory Diversity in the Extremely Acidophilic Acidithiobacillus genus: Insights into Anaerobic Metabolism

Pedro Sepúlveda[1,2], Carolina González[1,3], David S. Holmes[3], Jorge H. Valdés[2]

1. Centro de Genómica y Bioinformática (CGB), Facultad de Ciencias, Universidad Mayor (UM), Santiago, Chile
2. Centro de Bioinformática y Biología Integrativa (CBIB), Facultad de Ciencias de la Vida, Universidad Andrés Bello (UAB), Santiago, Chile
3. Center for Bioinformatics and Genome Biology (CBG), Fundación Ciencia y Vida (FCV), Santiago, Chile

Background:

Chemolithoautotrophic bacteria, especially members of the Acidithiobacillus genus, thrive in extremely acidic environments (pH < 3.5) and drive major biogeochemical cycles. They have developed several metabolic strategies to gather energy and nutrients using reduced inorganic compounds under aerobic and anaerobic conditions.

Results:

Using phylogenomics, comparative genomics, and in-silico regulatory reconstruction approaches, we generated a genome-wide assessment of metabolic and regulatory processes in 43 Acidithiobacillus representatives, identifying their conserved and variable metabolic pathways and regulatory instances for iron, sulfur, nitrogen, and hydrogen mobilization. We also identified conserved and specific pathways and their genomic features associated with anaerobic metabolism, including genes, their transcriptional units, and regulatory interactions, suggesting their potential role in metabolic speciation of these extreme acidophilic representatives.

Conclusions:

This genome-wide metabolic and regulatory reconstruction in Acidithiobacilli supports traceable connections between genomic divergence, metabolic diversity, and phenotypic variation, providing an integrated picture of ecophysiological roles of each species and its interactions under aerobic/anaerobic conditions in extremely acidic environments.

Cosechando rendimiento con inteligencia artificial

Felipe Gómez-Alvear[1], Belén Navarro[1], Maricarmen Osses[1], Juan Pablo Calderón[1], Romina Sepulveda[1], Ignacio Ramos-Tapia[1], Reinaldo Campos-Vargas[2], Daniel Aguayo-Villegas[1]

1. Molecular Biophysics and Bioinformatics Group, Center for Bioinformatics and Integrative Biology (CBIB), Facultad de Ciencias de la Vida, Universidad Andrés Bello, Santiago, Chile
2. Departamento de Producción Agrícola, Facultad de Ciencias Agronómicas, Centro de Estudios Postcosecha, Universidad de Chile, Santiago, Chile

Abstract:

El Smart Agro o agricultura 4.0 ha abierto nuevas herramientas e innovaciones que permite la toma de decisión oportuna por parte de los agricultores y empresarios agrícolas. La agricultura ya no depende únicamente del desarrollo de robots y máquinas más poderosas, sino también de la adquisición masiva y el análisis de datos en tiempo real y georeferenciado. Por ejemplo, la obtención temprana de datos certeros de producción permite mejorar la logística a través de planes de optimización de recursos basados en análisis de datos exhaustivos, entregándole mayor importancia a toda la cadena de valor, desde el campo hasta la mesa del consumidor. El análisis de estos datos agro climatológicos requiere de métodos Bioinformáticos que impulsan el mundo de las AgriTech. A modo de ejemplo del impacto de la AgroBioinformática presentamos el desarrollo de AgroIA, un prototipo basado inteligencia artificial y visión computacional que ocupa redes neuronales para estimar el rendimiento de producción de huertos de uva de mesa, con más de un 90% de certeza, apoyando al productor en una toma de decisión oportuna y efectiva.

Bayesian Probabilistic Inference and the Development of a Highly Multiplex long read sequencing protocol for SARS-CoV-2 genomes

García Labari I.[1], Ezpeleta J.[1,2], Casal P.E.[4], Posner V.[3,4], Villanova G.V.[3,4], Lavista-Llanos S.[1], Bulacio P.[1,2], Spetale F.E.[1,2], Murillo J.[1, 2], Angelone L.[1, 2], Paletta A.[5], Remes Lenicov F.[5], Cerri A.[6], Bolatti E.M.[4,6], Spinelli S.[7], Giri A.A.[4,6], Arranz S.[3], Tapia E.[1,2]

1. CIFASIS-UNR/CONICET
2. Facultad de Ciencias Exactas, Ingeniería y Agrimensura, UNR
3. Laboratorio Mixto de Biotecnología Acuática, Centro Científico Tecnológico y Educativo Acuario del Río Paraná
4. Facultad de Ciencias Bioquímicas y Farmacéuticas, UNR
5. INBIRS-UBA-CONICET
6. Grupo Virología Humana del IBR-CONICET/UNR
7. IDICER-CONICET/UNR

Background:
Viral genome sequencing allows identifying the evolutionary relationships among viruses, monitoring the validity of diagnostic tests, and investigating potential transmission chains. The objective of this work was the development of a complete protocol, from bench to bedside, for whole-genome, highly multiplexed SARS-CoV-2 sequencing. Towards this goal, we relied on a previously reported (in-silico) family of barcodes (NS-watermark) designed to deal with the high error-rates of long-read sequencing platforms. We used this protocol to identify the circulating variants and evolution of SARS-CoV-2 in Santa Fe, Argentina.

Results:
We built upon the amplicon tiling strategy described previously by Quick J. et al 2017 for the rapid whole-genome virus sequencing of clinical samples and coupled it with the NS-watermark multiplex sequencing strategy described previously by Ezpeleta J. et al. 2017. We focused on the SARS-CoV-2 multiplex-PCR 1.5 Kb amplification protocol (2x12-plex reactions), originally designed for the expensive and not portable PacBio sequencing machines, and adapted it for a low-cost and portable MinION alternative. We developed a multiplex sequencing protocol with barcoding sets of increasing size, 12, 48, and 96, out of a major set of 4096, and modified the multiplex-PCR protocol to allow double-end symmetrical-barcoding of amplicon samples with these rather long barcodes (36 nt). Pools of 12, 48, and 96 samples (including technical replicates) were sequenced together on the MinION sequencer. After base-calling and trimming of sequencing adapters, reads were individually deconvoluted using an approximate bayesian inference approach for the identification of individual barcodes (implemented by the NS-watermark decoding software). The use of a Bayesian inference approach allows a fine control of the critical trade-off between the rate of read recovery and the crosstalk-rate. Even for 96 samples, high coverage rates (> 98% of the genome) and depths (> 30X in each amplicon fragment of 1.5 Kb) were obtained. A subset of 110 complete genomes collected (March-December, 2020, available at GISAID) from individuals residing in 43 localities in the south of the province of Santa Fe were classified by dynamic lineage taxonomy with the Pangolin COVID-19 Lineage Assigner and were analyzed phylogenetically by IQ-TREE. The genomes obtained corresponded to 6 lineages, in coincidence with what was observed in other Argentine provinces during 2020. Clusters of sequences with close geographical proximity were identified, evidencing chains of viral transmission within the province of Santa Fe and / or with neighboring provinces.

Conclusions:
Our results validate the multiplex sequencing methodology developed with the NS-watermark barcodes that makes it possible to democratize genomic sequencing for the active surveillance of SARS-CoV-2 and may be extended to other emerging viruses in the future.

Uncovering the regulatory network that modules the transcriptional response to SARS-CoV-2 infection in humans

Mónica Padilla-Gálvez[1], Leo J. Arteaga-Vázquez[1], Karen J. Nuñez-Reza[1,2], Ana B. Villaseñor-Altamirano[1], Alejandra Medina-Rivera[1]

1. Laboratorio Internacional de Investigación sobre el Genoma Humano (LIIGH)
2. Instituto Nacional de Ciencias Médicas y Nutrición "Salvador Zubirán" (INCMNSZ)

Background:
The pathophysiology underlying COVID-19 across tissues and cell types upon SARS-CoV-2 infection still has gaps in our knowledge. To some extent, cellular processes have been described with special attention in IFN responses and utilization of ACE2 and TMPRSS2 proteins. But how the transcriptional machinery ensemble within cells can give us a better understanding of the disease progression. Which transcriptional programs are activated or repressed can be better understood via gene regulatory network reverse engineering?

Results:
Here, we make use of multiple publicly available transcriptional data across COVID-19 patient's lung autopsies and A549, NHBE and Calu-3 cell lines, recapitulate how they relate to bronchoalveolar lavage-derived immune cells from COVID-19 patients for further comparison and describe the transcriptional regulatory networks from each source generated using the SCENIC workflow.

Conclusions:
We described the regulatory networks underlying SARS-CoV-2 infection across tissues and cell lines. We could identify clear IFN response networks as expected.

Deciphering the relationship between epigenetic marks and Transcription Factor Binding

J. Sebastián Contreras-Riquelme[1], Alberto J.M. Martin[1]

1. Network Biology Lab., Centro de Genómica y Bioinformática, Universidad Mayor

Background:

Transcription Factors (TFs) bind to specific patterns in the DNA to influence cell fate. In eukaryotic cells, the DNA needs to be organized to allow or to impede the binding of the transcriptional machinery. Chromatin structure depends on the action of several proteins, among them, there are proteins that introduce chemical modifications (i.e., acetylation or methylation) in histone tails. Additionally, other proteins can modify the actual DNA, adding methyl groups usually to cytosines. A third regulatory level relays in the 3D chromatin structure, being dynamically regulated by loop-forming proteins such as CTCF, present in the base of ~90% human/mouse chromatin loops. Chromatin loops block or get closer in space regulatory elements, delimiting domains of homogeneous histone modifications. How the combinations of these factors are related to the binding of TFs remains poorly understood. To unravel this, we applied a Random Forest (RF) algorithm that aims to predict the activation state of TF binding sites (TFBSs) and by doing so, reports the relevance of each mechanism.

Results:

In this work we have used the epigenome of five human cell lines, testing several ways to develop the dataset such as the bin size for chromatin fragmentation and the extension of neighbors of the promoter. Additionally, we experiment by changing hyperparameters of the RF such as the number of trees and max depth that each tree could be extended. Our best model uses majority vote of the tree reaching a True Positive Rate (TPR) of 0.71, a False Positive Rate (FPR) of 0.05, and a Precision (P) of 0.829. Whilst using Youden's distance threshold we got a TPR of 0.764, a FPR of 0.09 and a P of 0.906. These models also show that H3K4m2/3 and H3K9ac at 1 kilobase up/downstream of the TFBS are the most important histone makers to define the status of the promoters, while the DNA accessibility increases its importance in a proximal region of the TFBS.

Conclusions:

The results have shown that the combinations of epigenetic modifications near to the promoter site are effective predictors of TFBS activity, whilst the effect of more distant elements needs to be studied in depth.

CTCF binding prediction using genomic and epigenomic features

Camilo Villaman[1], Alberto J Martin[1], Mauricio Saez[1]

1. Network Biology Laboratory (NBL), Universidad Mayor. Chromatin, Epigenetics and Neurosciences Laboratory (CEN LAB), Universidad Mayor

Background:

CTCF is the most relevant insulator protein in vertebrates and it is involved in the establishment and maintenance of topologically associated domains (TAD), self-interacting domains with a higher interaction ratio of elements inside the domain than outside it. CTCF also is capable of forming CTCF loops, which are TADs flanked by convergent CTCF binding sites anchored by CTCF and Cohesin. Aberrant CTCF binding leads to a loss of TAD boundaries and a disrupted transcriptional landscape due to abnormal interactions inside and outside the TAD, and it has been linked with several diseases like neuropathies and cancer. To explore the role of CTCF in disease, we built a predictor of the binding state of CTCF binding sites (CBS) to assess the binding state of CTCF in different conditions.

Results:

In this work we present a new predictor based on Random Forests (RF) that uses genomic and epigenomic features, to predict the binding state of each CBS in the genome. The predictor was trained using four cell lines (K562, HeLa, MCF7, SK-N-SH) from the publically available reference epigenomes data from ENCODE. The predictor is currently on alpha, and it can predict with over 0.8 precision and recall in all of the four tested cell lines. The RF also reports the relevance of each feature in the CBS state prediction, determining that DNA accessibility is the most relevant factor to determine CBS binding, followed by DNA methylation.

Conclusion:

We built a CTCF binding predictor to evaluate the binding state of CBS using genomic and epigenomic features, and while the idea was to build a CTCF-specific binding predictor, many of the concepts mentioned could be applied to other DNA-binding proteins. The results of this predictor will be used to build a CTCF loop predictor to assess differential CTCF loop landscapes in different diseases.

Imbalanced Fetal State Classification from Cardiotocograms

Néstor Diaz[1], Viviana Peña[1], Tomas Escobar[1]

1. Grupo de Investigación en Inteligencia Computacional – GICO, Facultad de Ingeniería Electrónica y Telecomunicaciones, Universidad del Cauca

Background:
A cardiotocogram is a valuable tool for fetal monitoring because it can help physicians to assess fetal well-being. In addition, by identifying fetuses with potential risks, it will be helpful to define interventions that enhance outcomes. Artificial intelligence methods have proven beneficial in fetal state classification based on cardiotocogram monitoring. However, it is widespread to overlook the impact of class imbalance in the training process, and therefore the issues derived from oversampled classes are unknown. We approach the problem of multiclass imbalanced fetal state classification by using oversampling and subsampling for classes and synthetic data generation. We trained several models using well-known machine learning algorithms for prediction performance comparison, such as support vector machines, random forest, decision trees, and k-neighbors.

Results:
The dataset used in our study is publicly available from the UCI repository[1] Expert obstetricians classified data instances in three classes: normal (78%, 1655 instances), suspect (14%, 295 instances), and pathological (8%, 176 instances). We used the holdout strategy to validate the machine learning models with 70% for training/validation and 30% for testing. We created training balanced datasets, using three classes' distributions: equal proportion (same number of instances per class, 33.3% each); medium balanced (41% Normal, 33% Suspect, and 26% Pathological); lightly balanced (41% Normal, 33% Suspect, and 26% Pathological). We used two balancing strategies for each class distribution: random subsampling for majority class with oversampling for minority classes and Synthetic Minority Oversampling Technique (SMOTE). We assessed all the trained models were using the holdout testing set for unbiased assessing. A support vector machine achieved the best result for the original imbalanced data with radial basis function kernel (accuracy 0.861). From the trained models, the best results were obtained by two random forests trained with minority classes random oversampling on a medium balanced dataset (overall accuracy was 0.939) and with a dataset with lightly balanced classes (overall accuracy was 0.948). Both random forests excel over one of the best classifiers known and evaluated in similar conditions[2], which achieved an overall accuracy of 0.936, and F-scores of 0.97, 0.81, and 0.876 for the normal, suspect, and pathological classes, respectively. Our best model achieved F-scores of 0.964, 0.812, and 0.93 for the normal, suspect, and pathological classes, respectively.

Conclusions:
Our experiments showed that by selecting an appropriate strategy of balancing for the minority classes, the classification results for the suspect and pathological cardiotocograms are improved. However, in our experiments for equally balanced data, the improvements over the original dataset were negligible. As each data distribution used in our experiments showed distinct results, it is convenient to explore a set of models to use in an ensemble that exploits the advantages for each selected predictor.

[1] https://archive.ics.uci.edu/ml/datasets/cardiotocography

[2] http://koreascience.or.kr/article/JAKO201508160153714.page

Using current bioinformatics tools to analyze single-cell transcriptomic data

Enrique De La Rosa[1], Elena Sanchez-Luis[1], José Manuel Sánchez-Santos[1], Javier De Las Rivas[1]

1. Bioinformatics and Functional Genomics Group, Cancer Research Center (CiC-IBMCC, CSIC/USAL), Consejo Superior de Investigaciones Científicas (CSIC) & University of Salamanca (USAL), Salamanca, Spain

Background:

In recent years, breakthroughs in the development of next-generation sequencing (NGS) and high-throughput omic technologies have generated a deeper knowledge about our understanding of complex biological systems, human diversity and diseases. Many of these new genomic, transcriptomic, and other multiomic technologies are directed at the study of individual cells. These single-cell tools can uncover complex and unknown populations, including cellular heterogeneity, revealing regulatory relationships between genes and showing evolutionary relationships between different cell types. All these new single-cell technologies can be applied in oncology, neurology, immunology, microbiology and many other scientific areas, and are therefore an important tool for current biomedical research. One of these technologies is single-cell RNA-sequencing (scRNA-seq) that analyses the transcriptome of each cell and it is the one addressed in this study.

Results:

In this work, we evaluate and describe the principal methods and algorithms employed to analyze a single-cell transcriptomic object. scRNA-seq has two main parts: (i) sample preparation and experimental data production and (ii) bioinformatic computational analysis. Despite the part treated in this work is the second one, it is important to understand the origin of the object that we are working with. Apart from the explanation of the methods for the study of single-cell transcriptomic objects, we present some other important and specific tools for cell trajectory analysis and multimodal single-cell studies. For this last purpose, we show an example of a study on haematological cells that we have carried out.

Conclusions:

Our analyses of single-cell transcriptomic objects are based on the use and application of scran and Seurat algorithms. Some other specific methods that we apply are the cell trajectory analysis for which we used TSCAN functions; and we also use tSNE and UMAP procedures for the cell data visualization. The combination of all these computational methods in the analysis of complex single-cell data provides a clear insight into the biomolecular markers of individual human cells.

Mixture deconvolution methods of transcriptomic data to identify cellular composition and markers

N. Alonso-Moreda[1], A. Berral-Gonzalez[1], J.M. Sanchez-Santos[1], J. De Las Rivas[1]

1. Bioinformatics and Functional Genomics Group, Cancer Research Center (CiC-IBMCC, CSIC/USAL), Consejo Superior de Investigaciones Científicas (CSIC) & University of Salamanca (USAL), Salamanca, Spain

Background:

The study of variability between cell populations makes it possible to identify the activity of specific genes and to determine how changes in these cells affect certain developmental processes in organisms (morphogenesis, embryogenesis or cell differentiation) and the appearance or development of some diseases, such as cancer. The experimental techniques have some limitations associated for example to the limited knowledge of phenotypic and cell specific markers and the complexity of biological samples that include multiple cell types. In recent decades, computational techniques have been developed to solve the problem of heterogeneity and complexity in cell samples, known as deconvolution methods, designed to decompose a mixture consisting different cell types into their component elements. In this way, these methodologies make it possible to calculate cell proportions in the mixture and identify specific biomarkers.

Results:

In this work, two supervised deconvolution algorithms (DECONICA and LINSEED) have been implemented, which are capable of calculating cell proportions and identifying biomarkers, and three supervised algorithms (CIBERSORT, FARDEEP and ABIS) have also been applied. The latter require prior knowledge of marker genes and therefore can only calculate cell ratios. The main objective of the study is to carry out a comparative analysis of these five deconvolution methods. To select the most accurate method, we consider three perspectives: (1) The accuracy of the methods in the inference of the proportions of cell types using microarrays signal expression; (2) The precision of the methods that use RNA-Seq signal expression; (3) The accuracy of the methods identifying gene signatures.

Conclusions:

The results inferred by DECONICA showed cell distributions different from the known data, as well as no variability between samples. LINSEED showed accurate results, but it can only handle a small number of genes and, as for ABIS, the actual proportions are necessary to be able to apply the method correctly. In general, the best methods are CIBERSORT and FARDEEP, but, as they are supervised methods, they require prior knowledge of the expression of the genes included in the gene signature they use.

Articulación de estrategias “Ligand Based”, “Structure Based Drug Design” y de Inteligencia Artificial aplicadas al diseño innovador de compuestos bioactivos

Margot Paulino[1]

1. Centro de Bioinformática – DETEMA – Facultad de Química – Udelar, Uruguay

Abstract:

Nuestra investigación tiene como objetivo el desarrollo de conocimiento y formación de RRHH en la que se consolidan equipos de investigación multidisciplinarios (química orgánica y computacional, biología y bioquímica molecular, bioinformática, parasitología) y multisectoriales (universidades, institutos universitarios y empresas) con el objetivo ulterior de llamar la atención e interés de la industria para el desarrollo de tripanosomicidas efectivos, anticancerígenos (colonorectal y leucemia infantil) y nutracéuticos (antioxidación, antiinflamación, neuroprotección). Se presenta a manera de ejemplo el desarrollo de tripanosomicidas, medicamentos para enfermedades “huérfanas”, un tema desatendido por la industria farmacéutica, situación que se espera revertir planteando el desafío de articular herramientas potentes de investigación básica y multidisciplinaria que converjan en la candidatura de tripanosomicidas de baja toxicidad. El prometedor andamio molecular (1,4-ariloxi ((nafto/furan/tiazo/oxazolo) quinonas), promovió la síntesis de un conjunto “test” y una primer medida de inhibición del crecimiento de T.cruzi(epimastigotes y tripomastigotes) y citotoxicidad y, mediante anclaje reverso, una primer selección de posibles blancos farmacológicos: tripanotiona reductasa (TR), dihidrolipoamida deshidrogenasa(DHLDH) y glucosa-6-fosfato deshidrogenasa (G6PDDH). Incrementando la complejidad de los modelos experimentales e investigando una posible vinculación al estrés oxidativo, se describen las diferentes Fases implementadas para el aporte de diseños innovadores. Fase (I): se ha amplificado la evaluación del “test” con medidas parasitológica en amastigotes, enzimática (TR/DHLDH/G6PDH y sus contrapartes mamíferas) y de interferencia en el metabolismo de disulfuros de bajo peso molecular, medido mediante sondas redox. Un posible mecanismo no mediado por blancos se investiga por mecánica cuántica (DFT), midiendo energías libres e índices electrónicos en la formación de semiquinonas promotoras de especies reactivas de oxígeno (ROS). Fase (II): se han racionalizado los efectos medidos mediante una batería de procesos in-silico Ligand Based Drug Design, en la que realizamos medidas cualitativas y cuantitativas de la relación estructura-actividad y el diseño de farmacóforos de consenso que permiten bucear y recolectar estructuras en bases de datos masivas. Los blancos farmacológicos y sus posibles ligandos se buscan, diseñan y sus interacciones de observan y miden mediante las técnicas de Structure Based Drug Design (anclaje reverso, anclaje directo, modelado por homología e in-silico, y dinámica molecular). Fase III: con el desarrollado un algoritmo de búsqueda masiva operativo sobre bases de datos (BindingDatabase, Reaxys) en uso de herramientas de Inteligencia Artificial (Deep Learning y Reinforcing Learning), se ha completado el desarrollo para, finalmente, proponer nuevos quimio tipos que se incorporarán al ciclo de validación experimental.

Representación numérica de ligandos y sitios de anclaje para el diseño de modelos basados en inteligencia artificial

Pablo C. Garcia Briosso[1]

1. Centro de Bioinformática – DETEMA – Facultad de Química – Udelar, Uruguay

Abstract:

Representación numérica de ligandos y sitios de anclaje en estructuras tridimensionales de proteínas para el diseño y desarrollo de modelos basados en inteligencia artificial y redes neuronales profundas (Deep Neural Networks).

Differentially expressed intronless genes across cancer highlight their functional role in epigenetics

Maribel Hernández-Rosales[1], Gabriel E. Herrera-Oropeza[2], José A. Ramírez-Rafael[1], Guillermo Romero Tecua[3], Katia Aviña Padilla[1]

1. Cinvestav-Irapuato, Mexico
2. King's College London
3. Data-Pop Alliance

Background:

Eukaryotic genomes are mainly composed of a genetic structure that comprises a combination of exons generally interrupted by intragenic noncoding DNA regions termed introns. However, intronless genes (IGs), single-exon genes that lack introns presence, also co-exist in complex organisms. Notably, the abundance of introns in most genes of multicellular organisms entails regulatory processes associated with multiple splice variants missing in IGs. IGs exempting splicing events entail a higher transcriptional fidelity, making them potential biomarkers and targets for therapy that deserves careful consideration. Cancer is a complex disease that relies on progressive uncontrolled cell division linked with multiple dysfunctional biological processes. Tumor heterogeneity remains the most challenging feature in cancer diagnosis and treatment. From a genetic perspective, initiation and progression of tumorigenesis are related to genetic and epigenetic genome alterations. Aware of IG's potential clinical relevance, the present study aims to identify their functional role and expression and epigenetic marks across cancer.

Results:

We identified 940 protein-coding IGs in the human genome, being ~65% of them deregulated across eight analyzed cancer types. Our results show that differentially expressed IGs are highly shared among tumors and are abundantly involved in nucleosome and chromatin condensation as well as immune responses. Moreover, IGs tend to have a more induced expression when compared to MEGs. Notably, the upregulated IGs across cancer types are highly conserved HDACs deacetylate histones involved in gene regulation. Strikingly, methylated IGs show distinct epigenetic hallmarks, mainly promoters hypomethylation.

Conclusions:

This study highlights that differentially expressed human intronless genes across cancer types are prevalent in epigenetic roles. Their role is linked to the high conserved induction of histones in all the studied tumors. Particularly HDACs that regulate expression undergoing protein deacetylation. IGs induction may be the outcome of characteristic epigenetic hallmarks such as hypomethylation.

Eukaryotic intronless genes: insights into their functional significance and evolution

Aviña-Padilla Katia[1], Ramírez-Rafael José Antonio[1], Herrera-Oropeza Emilio[2], Varela Echavarría Alfredo[3], Hernández-Rosales Maribel[1]

1. Cinvestav-Irapuato, México
2. King's College London
3. Instituto de Neurobiología UNAM

Background:
The structure of eukaryotic genes is generally a combination of exons interrupted by introns removed by RNA splicing to generate the mature mRNA. However, a subset of genes comprises a single coding exon with introns in their untranslated regions or are intronless genes (IGs), lacking introns entirely. The latter code for essential proteins involved in development, growth, and cell proliferation and their expression has been proposed to be highly specialized for neuro-specific functions and linked to cancer, neuropathies, and developmental disorders. The abundant presence of introns in eukaryotic genomes is pivotal for the precise control of gene expression. Notwithstanding, IGs exempting splicing events entail a higher transcriptional fidelity, making them even more valuable for regulatory roles.

Results:
This work aimed to infer the functional role and evolutionary history of IGs centered on the mouse genome. We identified 1,116 IG functional proteins validating their differential expression in transcriptomic data of embryonic mouse telencephalon. Our results showed that overall expression levels of IGs are lower than those of MEGs. However, strongly up-regulated IGs include transcription factors (TFs) such as the class 3 of POU (HMG Box), Neurog1, Olig1, and BHLHe22, BHLHe23, among other essential genes including the β-cluster of protocadherins. Striking was the finding that IG-encoded BHLH TFs fit the criteria to be classified as microproteins. Finally, predicted protein orthologs in other six genomes confirmed high conservation of IGs associated with regulating neural processes and with chromatin organization and epigenetic regulation in Vertebrata.

Conclusions:
This study highlights that IGs are essential modulators of regulatory processes, such as the Wnt signaling pathway and biological processes as pivotal as sensory organ developing at a transcriptional and post-translational level. Overall, our results suggest that IG proteins have specialized, prevalent, and unique biological roles and that functional divergence between IGs and MEGs is likely to be the result of specific evolutionary constraints.

Study of stratification based on transcriptomic in subjects with Alzheimer’s disease using individualized co-expression networks

Verónica Latapiat[1, 2], Alberto JM Martín[2], Mauricio Saez[3], Inti Pedroso

1. Programa de Doctorado en Genómica Integrativa, Vicerrectoría de investigación, Universidad Mayor, Santiago, Chile
2. Laboratorio de Biología de Redes, Centro de Genómica y Bioinformática, Facultad de Ciencias, Universidad Mayor, Santiago, Chile
3. Chromatin Epigenetics and Neuroscience, Centro de Genómica y Bioinformática, Facultad de Ciencias, Universidad Mayor, Santiago, Chile

Abstract:

Alzheimer's disease is the most prevalent form of dementia and increasing public health concern in aging societies. An essential factor in the difficulty and failure of treatments for this disease is associated with the assumption that Alzheimer’s disease patients are a homogeneous group, but this fact can hide subgroups with potential differential sensitivity to therapies. In genomic, system malfunction is usually studied by the correlation between the expression of pairs of genes in many samples as coexpression networks, but the traditional approach displays averages, erasing the heterogeneity of each individual. However, personalized coexpression networks allow identifying gene associations for each individual. Our objective is to identify differential modules in non-cognitive impairment and Alzheimer’s disease that can allow the stratification of subjects using individualized coexpression networks. We used a large publicly available dataset, later, we applied the WGCNA package and individualized methods to know differentiated biological processes between diagnostics groups and Alzheimer’s disease diagnosis. Our results have shown that individualized networks detect differences in biological data sets with known classifications. Using spectral clustering strategies we found disease-related modules that differ in terms of module numbers per sample and their number of genes using individualized coexpression networks. We can measure the preservation among modules of different samples and correlate with clinical metadata. These findings promise to identify specific changes in molecular interactions for individuals with Alzheimer's disease based on modules and contribute to integrating different approaches in the study of heterogeneity in conditions on a genomic level of other complex diseases.

Analysis of transcriptional changes associated with CRE elements in a murine model of Huntington's disease due to the interaction of CBP with the mutant Huntingtin protein

Sandra Arancibia-Opazo[1,2,3], Alberto J.M. Martin[2], Mauricio Sáez[3]

1. Programa de Doctorado en Genómica Integrativa,
Universidad Mayor, Chile
2. Laboratorio de Biología de Redes. Centro de Genómica y Bioinformática, Facultad de Ciencias
Universidad Mayor, Chile
3. CENLab. Centro de Genómica y Bioinformática, Facultad de Ciencias,
Universidad Mayor, Chile

Background:

Huntington's disease (HD) is a neurodegenerative disorder caused by an abnormal expansion in the number of CAG trinucleotide repeats within the HTT gene. Several studies have shown that the mutant huntingtin protein (mHtt) captures different proteins to form a nuclear aggregate, causing for example, an association with several transcriptional regulators such as the CREB-binding protein (CBP). CBP captured by mHtt has an altered function that causes altered acetylation patterns in histones presented in neurons, cellular toxicity and dysregulation of the cAMP response element binding protein (CREB). CREB regulates several neuroprotective processes and could play an important role in HD, so given this, this protein could cause a decrease in its transcriptional activity and therefore decrease the activity of many CRE elements.

Results:

Through the extraction of striatum tissue samples of wt and R6/2 mice, and a subsequent RNA sequencing, we generated transcriptomic information related to early stages of the disease. This information has allowed us to identify relevant transcriptional changes, that through the progression of the disease, there is a progressive increase in differentially expressed genes (DEG). Subsequently, in order to modelate how these changes have been carried out, we developed a bioinformatic method to search putative genes regulated by CRE. Using this application, we found various DEG genes that we found to be linked to CRE regions, including some of them that have been described in literature. Finally, through the development of Gene Regulatory Networks for both status, WT and illness model, we have modelled how transcriptional regulation is carried out in both conditions, comparing them to gathering information about key regulatory changes in early stages of HD given the decrease in CBP and the subsequent misregulation in CRE cascade.

Conclusions:

The decrease levels in genes close to CRE is due to the possible decrease in CBP levels. Affecting the CREB-CRE function, having a lower activation of genes. The gene regulatory network provides information for further analysis, in order to determine the key genes and the essential TFs in HD.

Actas del Taller de la Red Iberoamericana RIABIO sobre Inteligencia Artificial aplicada a BioData. nº 1, 2021.
Libro digital - PDF

Paginas tipeadas provistas por autores

RiaBio Editora
Campus Miguel de Unamuno s/n, 37007 Salamanca, España
coreRiaBio@cifasis-conicet.gov.ar
2021