virtual

10-12 Nov 2021

01001
00100
10100

# XI CAB2C

Congreso Argentino de Bioinformática y Biología Computacional

Abstract Book

http://bit.ly/cab2c-2021

a2b2c

bioinformatica_ar

#XICAB2C

10100
10011

Druggability assessment algorithm based on Composition, Transition and Distribution descriptors and publicly available predictive tools

Juan Ignacio Alice[1,2], Santiago Rodríguez[1,2], Lucas N. Alberca[1,2], Carolina Bellera[1,2], Alan Talevi[1,2]

[1] LIDeB- Laboratorio de Investigación y Desarrollo de Bioactivos (Facultad de Ciencias Exactas UNLP, La Plata, Argentina), [2] CONICET Consejo Nacional de Ciencia y Tecnología, Argentina.

srodriguez@biol.unlp.edu.ar; jalice@biol.unlp.edu.ar

Background:

In the framework target-guided drug discovery it is important to be able to assess the druggability of the proposed drug target prior to the implementation of a drug discovery project. Druggability is a concept coined by Hopkins and Groom to refer to the ability of a protein to be modulated by small, drug-like molecules.

Results:

A dataset of 222 proteins druggable and undruggable was compiled, and it was split into a training set for model building and an independent test set for model validation. The training set was then used to infer linear classifiers capable of prospectively discriminating druggable from non-druggable targets. Two algorithms were built and validated for such task. The first one uses CTD (Composition, Transition and Distribution) descriptors, while the second combines CTD descriptors with already reported and validated online druggability assessment tools. 14 druggability predictors were derived from online tools and 147 CTD descriptors were computed using the PyProtein module from PyBioMed library. Using a combination of feature bagging and forward stepwise feature selection, 1000 linear models were built using either a combination of online tools plus CTD or CTD descriptors alone .

The best individual model for CTD descriptors displayed an accuracy of 0.803, a precision of 0.738 and recall of 0.939 on the test set, while the best individual model emerging from the combination of CTD descriptors and online tools showed an accuracy of 0.871, a precision of 0.800 and a recall of 0.848.

Conclusions:

Any target-focused, rational drug discovery initiative starts with the choice and validation of an adequate drug target. Drug target validation implies, among other studies, guaranteeing that the chosen target is druggable. Here, we have reported an algorithm based on CTD descriptors and druggability descriptors derived from online tools, capable of differentiating, with remarkable accuracy druggable from non-druggable proteins in a fast and cost-efficient manner.

Oral Presentation

OBI: A computational tool for the analysis and systematization of the positive selection in proteins

Julián H. Calvento[1], Franco Leonardo Bulgarelli[2], Ana Julia Velez Rueda[1]

1. Departamento de Ciencia y Tecnología, CONICET, Universidad Nacional de Quilmes.
2. Mumuki.org

Positive selection analysis is a bioinformatic prediction technique with multiple applications, including, for example, vaccine design or the detection of new drug-resistant pathogenic variants (Chen et al., 2004; Duvvuri et al., 2009). However, applying these analyses on a large scale requires automation and optimization in computing speed and interoperability between technological tools, which makes it hard to achieve. Here we present OBI, an open-source tool that facilitates the analysis of positive selection on a large scale. We have implemented a stand-alone command-line app, developed entirely in Python, that can be freely used and installed through Conda's distribution[1].

Our tool integrates different computer and bioinformatics tools optimized for the prediction of positive selection. Receiving only a protein fasta sequence, our tool retrieves the homologous proteins(Johnson et al., 2008),and gene sequences using Entrez (Maglott et al., 2005) and performs the positive selection analysis using Hyphy (Pond et al., 2005). Furthermore, OBI links the evolutionary information with the structural data available for the protein of interest, allowing the user the easy detection of positive selection cases related to structural changes and their possible link with the activity and function of proteins.

Our tool presents very significant contributions in the field of software development since it involves the design and implementation of a processing architecture for a problem of great interest to the field of bioinformatics. We hope to provide with OBI a tool that reliably speeds up the evolutionary and structural analysis of proteins on a large scale.

Bibliography

Chen, L., Perlina, A., and Lee, C.J. (2004). Positive selection detection in 40,000 human immunodeficiency virus (HIV) type 1 sequences automatically identifies drug resistance and positive fitness mutations in HIV protease and reverse transcriptase. J. Virol. *78*, 3722–3732.

Duvvuri, V.R.S.K., Duvvuri, B., Cuff, W.R., Wu, G.E., and Wu, J. (2009). Role of Positive Selection Pressure on the Evolution of H5N1 Hemagglutinin. Genomics, Proteomics & Bioinformatics *7*, 47–56.

Johnson, M., Zaretskaya, I., Raytselis, Y., Merezhuk, Y., McGinnis, S., and Madden, T.L. (2008). NCBI BLAST: a better web interface. Nucleic Acids Res. *36*, W5-9.

Maglott, D., Ostell, J., Pruitt, K.D., and Tatusova, T. (2005). Entrez Gene: gene-centered information at NCBI. Nucleic Acids Res. *33*, D54-8.

Pond, S.L.K., Frost, S.D.W., and Muse, S.V. (2005). HyPhy: hypothesis testing using phylogenies. Bioinformatics 21, 676–679.

[1] Obi v1.0.0: https://anaconda.org/jcalvento/obi
Obi code source: https://github.com/jcalvento/obi

Poster Session: http://bit.ly/cab2c-2021-posters
Poster Calvento et al: https://youtu.be/rFp0sSm2l9w

CardIAP: Calcium images analyzer web application

Ana Julia Velez Rueda[1], Agustín García Smith[1], Luis Alberto Gonano[2], Silvina Fornasari[1], Gustavo Parisi[1] and Leandro Matías Sommese[1]

1. Departamento de Ciencia y Tecnología, CONICET, Universidad Nacional de Quilmes.
2. Centro de Investigaciones Cardiovasculares, CONICET, Facultad de Medicina, UNLP, Argentina.

Motivation

Cardiovascular diseases (CVD) are the leading cause of global death (GBD, 2018), and in particular, arrhythmias represent a major portion of these deaths (~15–20 %) (Srinivasan and Schilling, 2018). Since abnormal Ca2+ handling is linked to arrhythmias, understanding calcium (Ca2+) management is one of the cardiovascular research's main goals (Bers, 2014; Landstrom et al., 2017). The use of fluorometric techniques in isolated cells, loaded with Ca2+ sensitive fluorescent probes allows the quantitative measurement of dynamic events that occur in living, functioning cells. The Cardiomyocytes Images Analyzer Application (CardIAP) covers the need for tools to analyze and retrieve information from confocal microscopy images, in a systematic, accurate, and fast way.

Results

Here we present the CardIAP web app, an automated method for the identification of spatio-temporal patterns in a calcium fluorescence imaging sequence. Through this tool, users can analyze single or multiple Ca2+ transients from confocal line-scan images and obtain quantitative information on the dynamic response of the stimulated myocyte.

Our web application also allows the user the extraction of data on calcium dynamics in downloadable tables and plots, simplifying the calculation of the alternation and discordance indices and their classification. CardIAP could assist in studying the underlying mechanisms of anomalous calcium release phenomena.

Availability and implementation

CardIAP is an open-source app, entirely developed in Python, which can be freely accessed and used at http://cardiap.herokuapp.com/.

Poster Session: http://bit.ly/cab2c-2021-posters
Poster Velez Rueda et al: https://youtu.be/CZ_O5Hm2bjU

From evolution to folding of repeat proteins

Ezequiel Galpern[1,2], Diego Ferreiro[1,2]

[1]Departamento de Química Biológica, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Argentina. [2] Instituto de Química Biológica de la Facultad de Ciencias Exactas y Naturales (IQUIBICEN), CONICET, Argentina

Background:

Repeat proteins are made with tandem copies of similar amino acid stretches that fold into elongated architectures. Due to its symmetries, they are a unique system to model how evolutionary constraints at the sequence level can impact tertiary structure, folding and function. Ankyrin repeat proteins are usually described as formed with linear arrays of tandem copies of a 33 residues length motif that folds to a alpha-loop-alpha−beta-hairpin/loop. Being one of the most common repeat proteins in nature, these molecules are believed to provide specific protein-protein interactions. Despite the most studied Ankyrins have few repeat units, array length distribution extends to proteins with more than 120 repeats.

Results:

In this work we departed from a curated Ankyrin 150000 sequence dataset. We combined an inverse Potts model scheme with an explicit mechanistic model of duplications and deletions of entire repeats for calculating the evolutionary parameters of the system. We used the evolutionary energy obtained to input a folding toy model for repeat proteins, a coarse-grained 1D ising model where each spin corresponds to a protein sequence fragment and has to be folded or unfolded. Monte Carlo simulations of the model allowed as to get thermal unfolding curves that are compatible with experimental ones. In addition, we made a large scale quantitative folding characterization including stability, presence or absence of intermediaries, free energy barriers and cooperativity.

Conclusions:

Folding key features related to protein function can be predicted from sequence, paving the way for guiding protein design with an evolutionary model.

Oral Presentation

Viral metagenomics analysis of six New World bats species from Argentina

Agustina Cerri[1], Elisa M. Bolatti[1,2], Gastón Viarengo[3], Tomaz M. Zorec[4], María E. Montani[5,6,7], Pablo E. Casal[2], Lea Hosnjak[4], Violeta Didomenica[6], Diego Chouhy[1,2,3], Rubén M. Barquez[6,7], Mario Poljak[4], Adriana A. Giri[1,2].

[1] Grupo Virología Humana, Instituto de Biología Molecular y Celular de Rosario (CONICET), Suipacha 590, Rosario 2000, Argentina.
[2] Área Virología, Facultad de Ciencias Bioquímicas y Farmacéuticas, Universidad Nacional de Rosario, Suipacha 531, Rosario 2000, Argentina.
[3] DETxMOL S.A, Centro Científico y Tecnológico de Rosario, Ocampo y Esmeralda, Rosario 2000, Argentina.
[4] Institute of Microbiology and Immunology, Faculty of Medicine, University of Ljubljana, Zaloška 4, SI-1000 Ljubljana, Slovenia.
[5] Museo Provincial de Ciencias Naturales "Dr. Ángel Gallardo", San Lorenzo 1949, Rosario 2000, Argentina;
[6] Programa de Conservación de los Murciélagos de Argentina, Miguel Lillo 251, San Miguel de Tucumán 4000, Argentina.
[7] Programa de Investigaciones de Biodiversidad Argentina, Facultad de Ciencias Naturales e Instituto Miguel Lillo, Universidad Nacional de Tucumán, Miguel Lillo 205, San Miguel de Tucumán 4000, Argentina.

Background:
Bats are considered one of the most important natural reservoirs of a variety of zoonotic viruses, many of which (e.g. SARS-CoV-2 coronavirus) cause severe human diseases. Therefore, characterizing viruses of bats inhabiting different geographical regions is important not only for understanding their viral diversity but also for detecting viral spillovers between animal species. Herein, viral diversity of six bat species from Argentina was investigated using a metagenomic approach.

Results:
Specifically, selected fecal samples of 29 individuals from six different bat species, inhabiting two different geographical sites, were prepared and pooled by species, sex and collection site. Subsequently, enriched viral DNA was sequenced on the Illumina MiSeq platform and the obtained reads were trimmed, filtered and cleaned using several bioinformatics approaches. Finally, the target sequences were subjected to the viral taxonomic classification. A total of 4,520,370 read pairs were sequestered from the enriched pooled samples and, after quality filtering and trimming procedures, the taxonomic classification approach revealed that 20% of sequences mapped to viral taxa. Sequences from *Parvoviridae, Circoviridae, Genomoviridae, Papillomaviridae, Herpesviridae*, *Poxviridae* and *Arteriviridae* were the most prevalent among vertebrate viral families and identified in all bats species included in this study. In comparison to fecal pooled samples collected at *T. brasiliensis* colony in Rosario, higher diversity of vertebrate viral families was identified in fecal pooled samples from bats inhabiting Villarino Park (individual bats).

Conclusions:
Our findings provide new insights on viruses present in different bats species of our region, living in close contact with humans, and contribute to the understanding of their possible role of pathogen reservoirs.

Poster Session: http://bit.ly/cab2c-2021-posters
Poster Cerri et al: https://youtu.be/CekTeeA2lOk

Comparative genomic analysis reveals why Pleuronectiformes have small genome size in comparison to other teleosts

Fernando Villarreal[a], Nicolás Stocchi[a], Jordi Viñas[b] and Alejandro S. Mechaly[c]

[a]Instituto de Investigaciones Biológicas (IIB-CONICET-UNMdP), Facultad de Ciencias Exactas y Naturales, Universidad Nacional de Mar del Plata, Mar del Plata, Argentina.
[b]Laboratori d'Ictiologia Genètica, Departament de Biologia, Universitat de Girona, Girona, Spain.
[c]Instituto de Investigaciones en Biodiversidad y Biotecnología (INBIOTEC-CONICET), Mar del Plata, Argentina.

Background:

The black flounder (*Paralichthys orbignyanus,* order Pleuronectiformes) is an economically important marine fish with aquaculture potential in Argentina. We have recently sequenced, assembled and annotated its genome and analysis of comparative genome architecture with other genomes of teleost was also performed. The black flounder belongs to the order Pleuronectiformes and along with pufferfishes, seahorses and pipefishes, and anabantoidei (the group including bettas and gouramis) seem to have which presents a smaller size than most other teleost groups. Genome size variation seems to be related from a balance between genome-level mechanisms that tend to increase the genome size (duplication, transposable elements and polyploidy) and those genetic mechanisms that tend to decrease genome size (deletions and DNA repair mechanisms). However, this explanation is probably simple and incomplete. Thus, the principal objective of this study was to determine if the order Pleuronectiformes size and particularly, the black flounder, has a small genome size, considering it as the c-value, compared to the other teleost orders and determine if the exon and intron size of all the genes in Pleuronectiformes are smaller than the homologous genes in the other teleost genomes contained in Ensembl.

Results:

We analyzed i) c-values obtained from databases and ii) 21 complete genomes from different fish taxonomic groups (including *P. orbignyanus* and 2 other Pleuronectidei species). Based on annotations, we extracted genome-wide gene features (exons and introns) and calculated the number of genes, transcript models, exons and introns. Regarding c-values, we confirmed that among fish, *P. orbignyanus* size is amongst one of the smallest registered for Pleuronectiformes, which in turn show a low average c-value, significantly higher only to Tetraodontiformes. We were surprised to find that the small genome size in *P. orbignyanus* may be associated with lower exon number per transcript, which reduces overall transcript size, and it is also correlated with smaller protein sizes. Distribution of exon size at genome level is conserved across species. The intron size is variable when comparing species. In *P. orbignyanus*, the amounts of very large introns ($>10^4$ bp) and very large exons ($>3x10^3$ bp) are very low.

Conclusions:

As a result, the study concluded that the reduced flounder's genome is associated with the reduced transcript size, mainly by a reduction in exon number, but also by a reduction in large introns. As a conclusion, both components appeared to be involved in the genome-reduction strategy of Pleuronectiformes. Nevertheless, the lower abundance of repetitive elements such as transposons and similar structures cannot fully ruled out as alternative/complementary hypothesis of reduction of genome size in this species.

Poster Session: http://bit.ly/cab2c-2021-posters
Poster Villarreal et al: https://youtu.be/bFotrBCJyBk

Application of Ensemble Learning Approaches to Identify Inhibitors of the Main Protease of SARS-CoV-2

Prada Gori D. N.[1], Alberca L. N. [1], Alice J. I. [1], Caram F. N. [1], Andrea Medeiros[2,3], Martín Fló[2,3], Virginia López[2,3], Santiago Ruatta[2,4], Marcelo Comini[2], Bellera C. L. [1], Talevi A. [1]

[1]Laboratorio de Investigación y Desarrollo de Bioactivos (LIDeB), Facultad de Ciencias Exactas, Universidad Nacional de La Plata, La Plata, Buenos Aires, Argentina. [2] Institut Pasteur de Montevideo, Montevideo, Uruguay. [3] Universidad de la República, Uruguay. [4] Universidad Nacional del Litoral, Santa Fe, Argentina.

Background:

The Main Protease (Mpro, also called 3CLpro) of SARS-CoV-2 is a druggable cysteine protease with a crucial role in processing CoV-encoded polyproteins which mediate the assembly of replication-transcription machinery, thus representing an excellent target for the development of antiviral compounds. In this work, data from 414 chemical diverse compounds previously tested against the MPro of SARS-CoV-2 (including 109 with in-house acquired data), were compiled and partitioned into representative training and test sets, employing an in-house recursive clustering method known as iRaPCA. Using *in house Python routines* combining feature bagging and forward stepwise feature selection, and conformation-independent molecular descriptors linear classifiers were generated. The best classifiers were subsequently combined in meta-classifiers and validated using retrospective screening experiments. At last, a prospective screening campaign was implemented.

Results:

The compiled molecules were divided into a balanced training set (with 80 active and 80 inactive) and a test set (consisting of 54 active compounds and 200 inactive compounds). This test set was split in two stratified subsets which were complemented with 1450 putative decoys each, to evaluate the performance of the models against in pilot virtual screening experiments. After generating 1000 individual linear models, the top 22 models with the best performance were combined using the MIN_SCORE operator, enhancing their predictivity and robustness. By analyzing Positive Predictive Value (PPV), surfaces, a cutoff value of 0.546 was chosen, associated with a specificity of 0.998 and a PPV value of 0.634 for a hypothetic yield of active compounds of 1%.

The ensemble of 22 models was applied in the virtual screening of different chemical libraries including DrugBank, DRH, NuBBe, as well as *in house* libraries. As an additional selection criterion, we evaluated if the hits also belonged to the applicability domain of the model, thus selecting 49 molecules, potential inhibitors of MPro.

Conclusions:

We generated a computational ligand-based model ensemble associated to excellent enrichment metrics in the retrospective screens; the ensemble is able to identify Mpro of SARS-CoV-2 inhibitors. After a prospective virtual screen, 49 in silico hits were selected, which are to be confirmed in vitro in the near future.

Poster Session: http://bit.ly/cab2c-2021-posters
Poster Prada Gordi et al: https://youtu.be/e8CJHnyb7mA

End-to-end deep model for pre-miRNA prediction

Jonathan Raad, Leandro A. Bugnon, Diego H. Milone and Georgina Stegmayer

Research Institute for Signals, Systems and Computational Intelligence sinc(i) (FICH-UNL/CONICET), Ciudad Universitaria, Santa Fe, Argentina.

Background:
MicroRNAs (miRNAs) are small RNA sequences with key roles in the regulation of gene expression at post-transcriptional level in different species. Accurate prediction of novel miRNAs is needed due to their importance in many biological processes and their associations with complicated diseases in humans. Many machine learning approaches were proposed in the last decade for this purpose, but requiring handcrafted features extraction in order to identify possible de novo miRNAs. More recently, the emergence of deep learning has allowed the automatic feature extraction, learning relevant representations by themselves. However, the state-of-art deep models require complex pre-processing of the input sequences and prediction of their secondary structure in order to reach an acceptable performance.

Results:
In this work we present the first full end-to-end deep learning model for pre-miRNA prediction. This model is based on Transformers, a novel neural architecture that uses attention mechanisms to infer global dependencies between inputs and outputs. It is capable of receiving the raw genome-wide data as input, without any pre-processing or feature extraction. The model has been validated through several experimental setups using the human genome, and it was compared with state-of-the-art algorithms obtaining 10 times better performance.

Conclusions:
The results showed that the model can identify, without any preprocessing, all the pre-miRNA sequences within a genome with very high precision and recall.

Poster Session: http://bit.ly/cab2c-2021-posters
Poster Raad et al: https://youtu.be/jjqfJyILo1c

RELATIVE RESIDENCE TIME ESTIMATION OF SMALL ALLOSTERIC MODULATORS

Exequiel E. Barrera*, Diego M. Bustos*

**Laboratorio de Integración de Señales Celulares. Instituto de Histología y Embriología Mendoza. CONICET, Argentina*

Background: Computational methods are vastly used in the search of drug candidates both in academia and the pharmacological industry. One of the principal criterium to select small compounds is by estimating their binding affinity to target macromolecules of interest. Common practice binding affinity calculations can be classified in a broad spectra of accuracy and computational costs. Starting from molecular docking, to more detailed molecular dynamics derived techniques such as MM/GBSA or Umbrella Sampling, up to the most accurate myriad of hybrid quantum QM/MM methods. A recent alternative to these thermodynamics approaches proposes changing the focus towards kinetic properties, searching specifically for compounds with longer residence times, that highly correlate with their efficacy and selectivity. In this work, we employed the tau Random Accelerated Molecular Dynamics (τRAMD). Briefly, this technique applies random forces over the center of mass of the ligand, accelerating its egress process. Averaged residence times of multiple simulated replicas highly correlate with experimental data. Our main goal is to evaluate the predictive behavior of the technique on different small compounds that could allosterically modulate 14-3-3, a protein that interacts with phosphorylated proteins regulating multiple cellular processes like apoptosis and cell differentiation, just to name a few of them.

Results: We performed atomistic MD simulations of the ε paralog of 14-3-3 in its apo-form and with norharmane bound to its allosteric site. Norharmane is an unsubstituted β-carboline and we selected it as our lead compound. In the presence of the ligand, 14-3-3 remained locked in a closed state, different from the apo system which fluctuated between opened and closed states. The described behavior was quantified by performing a solvent accessible surface area analysis of the amphipathic groove, that corresponds to the region where phosphorylated partners bind. For the relative residence time (τ) calculations we first evaluate forces of different magnitudes. Starting with 14 kcal.mol-1.Å-1, that resulted in extremely short τ. This could be explained by the shallow shape of the binding pocket. Forces of 2.4 kcal.mol-1.Å-1 were the more appropriate for our system, resulting in τ = 0.58 ± 0.2 ns. A set of substituted β-carbolines were later docked into the allosteric site and their relative residence times were measured with τRAMD. Seeking selectivity, we employed the same set of compounds to study on a different paralog, 14-3-3 γ.

Conclusions: We observed how norharmane can allosterically regulate the conformational behavior of 14-3-3, locking its amphipathic groove into a closed state. The τRAMD technique was fine-tuned for our specific system and substituted β-carbolines presenting the highest residence times for each paralog have been chosen to be examined in further experimental studies.

Oral Presentation

EFFECTS OF DIFFERENTIAL PROMOTER USAGE IN TRANSCRIPTIONAL NOISE GENOME-WIDE: A NEURONAL DIFFERENTIATION PROCESS ANALYSIS

Martín Iungman and Ignacio Esteban Schor
IFIBYNE (UBA-CONICET)
martin.iungman@gmail.com, ieschor@fbmc.fcen.uba.ar

Author keywords: Cell-to-cell heterogeneity, single cell RNA-seq, Promoter activity, Transcriptional noise

Background:
In the last decades it has come to light that “identical” cells, even in the same environment, are heterogeneous in terms of their gene expression patterns. At the RNA level, this variability between cells is mostly due to what we call transcriptional noise, which can be caused, among other factors, by the nature of transcription, a temporally discontinuous process occurring in irregular bursts. Due to previous reports highlighting the role of gene promoters on this process, we question whether, in genes with multiple alternative promoters, transcriptional noise varies when promoter usage changes, for example during cell differentiation.
To explore this relationship we decided to study differential promoter usage and transcriptional noise genome-wide in two different time-points of a human neuronal differentiation process in vitro. To identify the genes which have differential usage of alternative promoters between these time-points, we took data published CAGE (Cap Analysis of Gene Expression) from the FANTOM project (https://fantom.gsc.riken.jp/5/), a bulk RNA technique that allows the identification of the 5’ end of each transcript. We used a binomial generalized linear model to test changes in promoter usage between day 0 and 12 of the process. On the other hand, we analyzed single-cell RNA-seq data (Yao et al, Cell Stem Cell, 2017), using VarID method (Grün, Nat Methods, 2020), which allows for local noise quantification by defining homogenous cell groups with pruned KNN network. Corrected RNA level variation was used as a proxy for transcriptional noise, and differential noise between each pair of cell types was evaluated with a Wilcoxon test. Association between changes in transcriptional noise and promoter usage was assessed both with lineal models and Wilcoxon test.

Results:
We detected 663 and 73 genes with significant difference in promoter usage and transcriptional noise, respectively, between the two time-points. The results integrating these analyses suggest that there is not a strong association between these two phenomena.

Conclusions:
Our work presents an effort to integrate different genomic datasets to build and test hypotheses on how gene expression robustness is achieved in eukaryotic cells, and how relevant to this robustness is the presence of multiple promoters.
Submisson track: Genomics, transcrptomics and metagenomcs

Poster Session: http://bit.ly/cab2c-2021-posters
Poster Iungman et al: https://youtu.be/22SNGX0a3IQ

iRaPCA: A NOVEL METHOD FOR CLUSTERING OF SMALL MOLECULES.

Alberca Lucas[1,2,3], Prada Gori Denis[1,2], Bellera Carolina[1,2], Talevi Alan[1,2]

[1] LIDeB, Facultad de Ciencias Exactas UNLP, La Plata, Argentina,

[2]Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), CCT La Plata, La Plata, Buenos Aires, Argentina

[3]Laboratorio de Señalización y Mecanismos Adaptativos en Tripanosomátidos, Instituto de Investigaciones en Ingeniería Genética y Biología Molecular (INGEBI).

lucasalberca@gmail.com, alantalevi@gmail.com

Background:
Clustering of molecules implies the organization of a group of molecules into smaller subgroups (clusters) with similar features. Typical applications of this methodology involve the representative splitting of datasets for QSAR and the selection of representative in silico hits from in silico screening experiments for acquisition and submission to experimental confirmation. In this work, we present an in-house hierarchical representative sampling procedure for clustering of small molecules. The approach, which we called iRaPCA, is based on an iterative combination of the random subspace approach (feature bagging), Principal Component Analysis (PCA) and the k-means algorithm. Our method has been converted to webapp so that any user can upload their smiles and perform the clustering with their own parameters.

Results:
A new online tool for the clustering of molecules has been developed. We have tested our tool in 29 datasets containing between 100 and 5000 small molecules, while comparing these results with the clusters from three other well-known clustering methods (*Ward*, *Complete* and *Butina* methods), as a benchmarking exercise. In all cases, internal validation has been performed. The mean silhouette score obtained for the 29 datasets by our method was 0.9045, while from Ward, Complete and Butina methods the silhouette score was 0.43, 0.42 and 0.27, respectively. Regarding the number of clusters obtained and the percentage of outliers (atypical molecules), iRaPCA on average obtains, for the optimal clustering judging from the silhouette coefficient, 14.2 clusters and less than 1% of outliers, while the other methods on average generate, for the optimal clustering, 221.9 (Ward), 175.8 (Complete) and 127.6 (Butina) clusters and more than 10% of outliers (23% in the case of Butina).

Conclusions:
iRaPCA has shown a great potential for the generation of dense and separated clusters of molecules as have been demonstrated in the benchmarking exercise. The implementation of our method as a Web App allows users who are unfamiliar with programming to perform a quick and easy clustering of molecules using their own parameters.

Poster Session: http://bit.ly/cab2c-2021-posters
Poster Prada Gori Denis et al: https://youtu.be/A3o1lh7PJvQ

Pseudotemporal ordering of breast scRNA-seq data

Daniela Senra[1], Nara Guisoni[1], Luis Diambra[2]

1.Instituto de Investigaciones Físicoquímicas Teóricas y Aplicadas, CONICET, Universidad Nacional de La Plata.

2. Centro Regional de Estudios Genómicos, CONICET, Universidad Nacional de La Plata

Background: The human breast is an organ composed mainly of glandular, adipose and connective tissue. The basic structure of the mammary gland consists of lobular units that produce breast milk interconnected by an intricate system of ducts. The vast majority of breast tumors arise from the epithelial cells lining the terminal ductal lobular units. For this reason, the characterization of the healthy mammary epithelium is an important aspect to comprehend the origins of breast cancer. Therefore, it is of particular interest to understand the differentiation pathway of the mammary epithelium.

As single cell RNA sequencing use has widely increased, many trajectory inference techniques that use this data type have been developed in recent years. Most of the trajectory inference algorithms require the selection of a start cell to infer the path of differentiation, which is usually set by previously known stemness markers and may not be adequate due to dropouts in scRNA-seq.

We propose a Protein-Protein Interaction Network (PPIN) approach to identify the breast stem cells for later trajectory inference. We implement the method to calculate the activity of the PPIN for each cell in R. As the goal is to find the stem cells, the activity of the protein network associated to cellular differentiation process is a parameter that indicates the differentiation activity. In this way, the cells with the highest differentiation activity are determined and selected as the root for the trajectory.

Results: In this work, we analyzed publicly available scRNA-seq data from epithelial breast tissue. After preprocessing, dimensionality reduction, clustering and differential expression analysis three main cell types were identified and were annotated as Basal, Luminal immature (L1) and Luminal mature (L2).

We applied the diffusion maps dimensionality reduction technique, which considers data points (cells) as nodes of a graph and assumes a diffusion process on the graph. After the diffusion map embedding of the data, a pseudotime value can be assigned to each data point relative to the origin cell. We developed a method to determine the root of the differentiation process that does not depend on the previous knowledge of tissue specific stemness markers. A differentiation PPIN was constructed and employed to compute a differentiation activity score for each individual cell by using a custom-made algorithm. We applied this workflow to each person of the data-set separately and obtained consistent results of the root and the differentiation path.

Conclusions: Implementing a customized method to select the root based on a differentiation PPIN the differentiation trajectory of the epithelial cells of human mammary tissue was inferred using the technique of diffusion maps and pseudotime ordering. This pseudotime ordering allow us to reconstruct the gene expression profile during the differentiation process.

Oral Presentation

Performance evaluation of three machine learning algorithms for breast cancer classification

José Félix Rojas C.
Universidad Gerardo Barrios

Background:
Breast cancer is one of the main causes of death for women in the world. It can be detected by distinguishing malignant tumors, and detection generally depends exclusively on a specialist. Providing systems that allow automated tumor detection may be beneficial for healthcare systems, which require reliable diagnostic processes to recognize these types of tumors.
The objective of the study was to analyze the performance of three algorithms to identify breast cancer sets in the Wisconsin dataset. We used the following algorithms: k-nearest-neighbors, support vector machines, and random forest, as implemented in the caret package for R. Performance parameters of the algorithms and Receiver Operational Characteristic curves are presented.

Results:
The random forest algorithm had the best performance, with an accuracy of 0.9767, a kappa of 0.9469, a sensitivity of 1.000 and a specificity of 0.9662. The k-nearest neighbors algorithm had the second-best performance, with an accuracy of 0.9734, a kappa of 0.9388, a sensitivity of 0.9787, and a specificity of 0.9710. The support vector machines algorithm had an accuracy of 0.9666, a kappa of 0.9235, a sensitivity of 0.9681, and a specificity of 0.9662.

Conclusions:
The random forest algorithm performed best, though the other algorithms performed similarly.

Poster Session: http://bit.ly/cab2c-2021-posters
Poster Rojas: https://youtu.be/2hjxzhUBCh8

INTEGRATING GENE EXPRESSION PROFILES WITH NON-CODING SOMATIC MUTATION ANALYSIS TO IDENTIFY KEY REGULATORS OF METASTASIS IN TRIPLE-NEGATIVE BREAST CANCER

Pedro J. Salaberry[1], Camila D. Arcuschin[1] and Ignacio E. Schor[1,2]

[1] Instituto de Fisiología, Biología Molecular y Neurociencias (UBA-CONICET), Buenos Aires, Argentina

[2] Departamento de Fisiología, Biología Molecular y Celular, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Buenos Aires, Argentina.

Background:

Massive efforts for whole-genome sequencing of tumor samples have revealed the high frequency of somatic mutations in non-coding regions. However, limited effort has been made to understand the oncogenic potential of these mutations. Triple-negative breast cancer (TNBC) has the worst prognosis among breast cancer subtypes, with high rates of metastasis and lack of targeted treatments. In this work, we use gene expression data from isogenic TNBC cell lines with different metastatic potential to identify genes or pathways whose deregulation contributes to a more malignant phenotype in TNBC, while analyzing the frequency of potential regulatory mutations in their promoters.

Results:

We combined a known differential expression analysis method (DESeq2) and a BRCA regulatory network reverse-engineered from publicly available patient transcriptome data (aracne.networks) with an algorithm (VIPER) which allowed us to infer changes in the activity of specific regulators through the interrogation of the expression levels of their target genes. We analyzed all possible pairs of isogenic TNBC lines, clustered the resulting regulators according to their shared activity profile and functionally characterized each cluster by means of an overrepresentation analysis. This led us to a list of 344 differentially active regulators, many of which have already been linked to oncogenic processes. These were classified in 8 clusters with clear distinction in terms of overrepresented functions and pathways. In parallel, we identified active promoter regions in breast tissue using CAGE data from the FANTOM5 project, and matched those regions with reported TNBC mutations from COSMIC and PCAWG databases. Finally, we filtered those mutations based on their FATHMM-MKL functionality score and analyzed their distribution within the regulatory pathways previously described, identifying 25 high-score non-coding mutations harbored in active promoters of these sets of genes.

Conclusions:

This work presents an effort to understand the impact of regulatory mutations on tumor biology. We intend to expand the search for regulatory mutations to other direct and functional interactors of the found regulators. Additionally, we will evaluate the consequences of the reported regulatory mutations on the promoters' activity, in order to assess their relevance on the deregulation of gene regulatory networks involved in TNBC metastatic potential.

Poster Session: http://bit.ly/cab2c-2021-posters
Poster Salaberry et al: https://youtu.be/m0PXq0Nu_q8

FUSING INFORMATION SOURCES THROUGH CONVOLUTIONAL NEURAL NETWORKS FOR GENE REGULATORY NETWORK INFERENCE

Mariano Rubiolo, Matías Gerard, Leandro Vignolo

Research Institute for Signals, Systems and Computational Intelligence, sinc(i), FICH/UNL-CONICET,

Ciudad Universitaria, 3000 Santa Fe, Argentina

Background:

The inference of gene regulatory networks underlying a gene expression dataset has received special attention by scientists for understanding the role of each gene and their relationships in the biological processes at molecular levels. Several inference methods based on machine learning models have been proposed, considering biological information in some cases, with different performance levels. This work proposes to fuse the inferred networks by applying both Extreme Learning Machine and Genetic Algorithms models with the biological information available in Gene Ontology for the involved genes. Inspired by how a Convolutional Neural Network identifies a color image analyzing the typical three colors channel, this approach considers the reconstructed networks after running the models and its biological information as three information channels. To test this proposal, synthetic datasets were built using the *E. coli* and *Yeast* models provided by the GeneNetWeaver tool of DREAM challenge, and were used for obtaining the best hyperparameters configuration for the convolutional neural network. Several simulations of the same gene regulatory network were presented to the convolutional model in order to obtain a more accurate regulatory network between those genes.

Results:

The performance of the baseline ELM model (F1 = 0.636), is improved with application of the convolutional model over the obtained raw gene regulatory network, with an F1 of 0.682.

If the resultant network from the previous model is fused with the network inferred by the genetic algorithm model, the convolutional neural network improves their performance, scaling the F1 up to 0.741. And, if the biological information layer is added to the input, the convoluted network performs better than the previous, up to 0.846 for the F1 measure.

Conclusions:

Our experiments show that complementing machine learning models predictions with biological information of the genes involved in gene regulatory networks as the input of a convolutional model has demonstrated an improvement on the final identification of the networks underlying the gene expression data. Future work involves a deeper convolutional model hyperparameters evaluation, regarding not only model layers but also biological information at the input.

Poster Session: http://bit.ly/cab2c-2021-posters
Poster Rubiolo et al: https://youtu.be/e9_a_1tmQKE

Bioinformatic Identification of Nuclear Hormone Receptor DAF-12 in Plant Parasitic Nematode M. Incognita

Claudio David Schuster and Carlos Pablo Modenutti
IQUIBICEN-FCEyN-UBA
claudio.schuster.93@gmail.com, cpmode@gmail.com

Background:
The root-knot nematode (RKN) Meloidogyne spp. Is a plant-parasitic nematode responsible for infecting crops and producing losses of billions of dollars a year worldwide. These worms lay eggs on the plant root, and when they hatch they infect the root during the second juvenile stage J2. This J2 developmental stage is one of the most vulnerable, and therefore the best phase to develop a controlling strategy.
One of the most attractive candidates is DAF-12, a nuclear receptor present in C. elegans, that binds a family of compounds known as dafachronic acids (DAs). In favorable conditions, the protein DAF-9 is expressed in J2 stage, which completes the synthesis of DAs. DAs bind to DAF-12 and activate developmental genes, allowing the larva to become adult. Although several C. elegans DAF-12 homologues have been identified in genomes of different parasitic nematodes, attempts to identify it in the meloidogyne genus using Blast have not been successful. Given this background, in this study we searched the RKN Meloidogyne Incognita genome for orthologous of DAF-9 and DAF-12 using a strategy that uses Hidden Markov Models (HMMs) to filter the proteome to those proteins containing the necessary domains, and discards any possibility of paralogy by means of a phylogenetic analysis that includes validated sequences of the studied genes from other species. By combining these tools we developed a new strategy to find orthologous genes in genomes that allowed us to identify DAF-9 and DAF-12 in M. incognita.

Results:
We used the software HMMER (that uses HMMs) to filter possible DAF-9 and DAF-12 candidates from the M. incognita proteome, ensuring that they all contain the necessary domains. Next, we narrowed down the candidates using phylogenetic analysis by maximum likelihood (software phyML), including as references previously known sequences of these genes in other species like C. elegans. This allowed us to separate from all the candidates the possible orthologous from the possible paralogous genes. Phylogenetic analysis was performed both with protein sequences and protein-coding sequences. A total of 3 candidates for DAF-9 and 3 for DAF-12 were obtained.Next, we analyzed sequence similarity between candidates and reference sequences with pairwise alignments, which showed that M. incognita sequences are notoriously divergent from their counterparts in other species. Finally, we studied differential expression across life cycle stages of these 6 candidates using an RNAseq experiment from a previous study and compared the trends with their counterparts in C. elegans, being DAF-9 candidates the ones that most resemble in differential expression to that of C. elegans.

Conclusions:
These results strongly suggest that the 3 candidates for both DAF-9 and DAF-12 may be actual orthologous of those 2, as they all contain the necessary domains and fall in the same group as the reference sequences in the phylogenetic analysis. Moreover, a previous genomic study of M. incognita suggested that this parasite may be at least triploid, which is coherent with the finding of 3 candidates for each gene. Furthermore, experimental analysis corroborated these candidates.

Poster Session: http://bit.ly/cab2c-2021-posters
Poster Schuster et al: https://youtu.be/tWIo6j-So0k

Higher-order interactions in neural populations

Natali Guisande, Mariela Portesi, Fernando Montani

Instituto de Física La Plata (IFLP), CONICET - Universidad Nacional de La Plata, La Plata, Argentina

Background:

Advances in multi-unit recordings make possible the investigation of the simultaneous activity of a very large number of neurons. Previous research has proved that the activity of all neuronal inputs follows a nonGaussian distribution that shapes the population response. We have recently developed a dichotomized q-Gaussian population model of easy numerical implementation, that can capture the subtle changes of the input's heterogeneities. Within our approach, correlations across neurons arise from q-Gaussian inputs into threshold neurons, and higher-order correlations in the spiking output activity are quantified by the parameter q.

Results:

This work presents an exhaustive analysis of how input statistics are transformed in this threshold process into output statistics, by using a combination of tools of graph theory and information theory. It shows under which conditions higher-order correlations can lead to either bigger or smaller number of synchronized spikes in the neural population outputs.

Conclusions:

The results allow us to characterize the population firing activity obtained from simultaneous recordings of neurons across all layers of a simulated cortical microcolumn. We put special emphasis on how to perform computational estimations efficiently when considering higher-order interactions in the input connectivity. This enables us to precisely quantify the amount of correlations that are generated from higher-order interconnectivity of the common overlapping neuronal inputs.

Poster Session: http://bit.ly/cab2c-2021-posters
Poster Guisande et al: https://youtu.be/c7ciKnBTRUA

PARCE: Protocol for Amino acid Refinement through Computational Evolution

Rodrigo Ochoa[1], Miguel A. Soler[2], Alessandro Laio[3,4], Pilar Cossio[1,5]

1 Biophysics of Tropical Diseases, Max Planck Tandem Group, University of Antioquia, 050010 Medellin, Colombia
2 Italian Institute of Technology (IIT), Via Melen 83, B Block, 16152, Genova, Italy
3 International School for Advanced Studies (SISSA), Via Bonomea 265, I-34136 Trieste, Italy
4 The Abdus Salam International Centre for Theoretical Physics (ICTP), Strada Costiera 11, 34151 Trieste, Italy
5 Department of Theoretical Biophysics, Max Planck Institute of Biophysics, 60438 Frankfurt am Main, Germany

Background:

The in silico design of peptides and proteins as binders is useful for diagnosis and therapeutics due to their low adverse effects and major specificity. To select the most promising candidates, a key matter is to understand their interactions with protein targets. Here we present PARCE, an open source Protocol for Amino acid Refinement through Computational Evolution that implements an advanced and promising method for the design of peptides and proteins.

Results:

The protocol performs a random mutation in the binder sequence, then samples the bound conformations using molecular dynamics simulations, and evaluates the protein-protein interactions from multiple scoring. Finally, it accepts or rejects the mutation by applying a consensus criterion based on binding scores. The procedure is iterated with the aim to explore efficiently novel sequences with potential better affinities toward their targets.

Conclusions:

The protocol can be applied with any protein of interest to design bound peptides or proteins of biological interest. We provide a tutorial for running and reproducing the methodology. The code is available at: https://github.com/PARCE-project/PARCE-1

Poster Session: http://bit.ly/cab2c-2021-posters
Poster Ochoa et al: https://youtu.be/uGVdmMrWLyc

Machine learning applied to molecular identification of *Acinetobacter baumannii* Global Clone 1

Verónica Elizabeth Álvarez[1], María Paula Quiroga[1] and Daniela Centrón[1]

[1]Laboratorio de Investigaciones en Mecanismos de Resistencia a Antibióticos, Instituto de Investigaciones en Microbiología y Parasitología Médica, Facultad de Medicina, Universidad de Buenos Aires - Consejo Nacional de Investigaciones Científicas y Tecnológicas (IMPaM, UBA-CONICET), Buenos Aires, Argentina.

Background:

Acinetobacter baumannii is a Gram-negative opportunistic and nosocomial pathogen that causes a broad range of nosocomial infections. It has become a major global threat because of its high level of resistance to several families of antibiotics. The majority of *A. baumannii* isolates that are broadly resistant to antibiotics belong to two globally disseminated clones, known as global clones 1 (GC1) and 2. Over time, molecular methods with different degrees of resolution have been used to typing *A. baumannii*. With the increasing throughput and decreasing cost of DNA sequencing, whole genome sequencing (WGS) may be an alternative for rapidly identification of pathogens. Machine learning (ML) algorithms and statistics have been increasingly used to build models that correlate genomic variations with phenotypes and may help to predict bacterial phenotypes and genotypes.

Results:

In this work, we built predictive models of *A. baumannii* GC1 genome typing by training Support Vector Machine (SVM) and Set Cover Machine (SCM) ML classifiers. We split 500 *A. baumannii* genomes into short sequences (unitigs and k-mers) used as input to train and test the models. Model performance was evaluated in terms of sensitivity, specificity, accuracy, precision and F1 score. Once we obtained accurate classification models, we used another dataset composed by 4799 *A. baumannii* genomes to validate ML results. From SVM and SCM models, we inferred new putative genomic biomarkers for GC1 genomes not subjected to selective pressure such as antimicrobial resistance genes. Using SVM model predictions, we found a region of 367 nucleotides along *moaCB* gene that differs between the GC1 and non-GC1 genomes, becoming a biomarker that differentiates both groups. We proposed a molecular method for GC1 strain typing that can be applied to direct detection from clinical samples by PCR.

Conclusions:

In conclusion, we obtained a SVM model that made genotypic predictions based on the presence or absence of short genomic sequences present in *A. baumannii* GC1 and non-GC1 genomes and found a sequence biomarker that uniquely identifies CG1. Given the length of the biomarker found, it is well-suited for translation to the clinical settings to be used in typified methods such as PCR.

Poster Session: http://bit.ly/cab2c-2021-posters
Poster Alvarez et al: https://youtu.be/dWGubPtedko

GiRoS: A computer vision approach to estimate sunflowers performance

Flavio E. Spetale[1], Javier Murillo[1], Paolo Cacchiarelli[2], Gustavo Rodriguez[2], Elizabeth Tapia[1]

1- CIFASIS-UNR/CONICET. Centro Internacional Franco-Argentino de Ciencias de la Información y de Sistemas.
2- IICAR-UNR/CONICET. Instituto de Investigaciones de Ciencias Agrarias de Rosario.
spetale@cifasis-conicet.gov.ar

Background:

Sunflowers (*Helianthus annuus* L.) are one of the three most important crops in Argentina, along with soybeans and corn. Correct identification of the sunflower yields contribute to the efficient characterization of the techniques and products utilization that can be a reference for future planting processes. Sunflower yield is performed considering the difference between the number of achenes and the number of seeds, not all achenes contain seeds inside. Traditionally, this task is manually performed over crop samples. Besides being extremely time consuming, the accuracy tends to be poor. In this work, we face the problem of identifying and counting the number of achenes through a computer vision method. A web interface, the source code, test image datasets and a batch script is freely available at www.cifasis-conicet.gov.ar/bioinformatica/Girasol/.

Results:

The final estimation of sunflower yield in our proposal consists of four steps: i) detection of sunflower contour; ii) identification and counting of achenes; iii) identification and counting of seeds and iv) results report. In this work, we focus on the first two steps considering three useful flower development stages in the sunflower (R7 and R8, both post-anthesis reproductive stages) where achene colors change from white, to brown and finally to black. To tackle contour identification problem, a mask is generated using edge detection techniques with different vegetation indices, i.e. metrics generated to distinguish plant parts by means of three channels (R, G, B), that are then applied to the original image. The second step performs an individual achene identification considering two vegetation indices, area and a series of decision rules based on the color of the achene. The dataset to test our method has 50 white, 30 brown and 70 black real images with different characteristics such as shape, size, shadows, bright taken from a private field in Santa Fe - Argentina during the 2020 harvest where each jpeg image has 1280x720px resolucion.

Conclusions:

The general average accuracy over the dataset is 85%. In particular, the accuracy obtained for black, white and brown achenes was 85%, 89% and 81% respectively. We note that the intense bright and shadows from the border leaves in sunflower borders affect method precision. Without those effects, the proposed method performance may increase. Time reduces from hours to seconds for 10 images, moreover the result is objective, always providing the same result for the same photo, a situation which is not always true for two different persons performing the task.

Oral Presentation

The Synaptotagmin-1 C2B Domain and the stabilization of the fusion pore

Ary Lautaro Di Bartolo[1,2], Diego Masone[1,3]

1 Instituto de Histología y Embriología de Mendoza (IHEM) - Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Universidad Nacional de Cuyo (UNCuyo), 5500, Mendoza, Argentina.
2 Facultad de Ciencias Exactas y Naturales, Universidad Nacional de Cuyo (UNCuyo), 5500, Mendoza, Argentina.
3 Facultad de Ingeniería, Universidad Nacional de Cuyo (UNCuyo), 5500, Mendoza, Argentina.

E-mail: ldibartolo@mendoza-conicet.gob.ar

Background:
Predictions of the behavior of biological systems through computational simulations allow for the generation of valuable information that, combined with experimental data, improves scientific advance in biomedicine. The development of better models, more efficient algorithms and the impressive increase in computing power have allowed unprecedented scales and simulation times to be achieved.

In particular, lipid-lipid interactions in aqueous environments naturally form a surprising variety of self-assembling structures, among which lipid bilayers are a representative example given their important biological function and their use as a model in fundamental biophysical studies. The amphiphilic molecules of the bilayer are located with the hydrophilic groups oriented towards the interface with the water, keeping the nonpolar chains inside the membrane. In this way, the cell interior is isolated, although with selective permeability. This property is of crucial interest in biomedicine and biotechnology since it is this controlled permeability that is responsible for the transport of ions, molecules and nanoparticles through the membrane.

Although markedly more complex than the membrane pore, a fusion pore is an effective mechanism for connecting intracellular organelles and releasing vesicular contents during exocytosis. A complex rearrangement of lipid molecules takes place during the approximation of the membranes, their deformation and their fusion. Thermodynamically the process is unfavorable and it has been hypothesized that it is mediated by specialized proteins, such as SNARE and Synaptotagmine. The first step for a three-dimensional analysis of this process (of a large time scale in computational terms) requires a detailed description of the complex rearrangement of lipid molecules for the formation of a single fusion pore. Only after that, the energy cost of the process can be quantitatively evaluated, with and without the presence of synaptotagmine, and therefore compared.

Results:
After developing and implementing a collective variable (CV) that describes the conformational changes in the formation of the fusion pore, umbrella sampling, an enhanced sampling technique, is carried out to obtain the free energy profile by biasing the molecular dynamics (MD) simulations with GROMACS.
Following an event-oriented classification of the different stages of the fusion pore, we can say that there are three stages:

- Membrane fusion.
- Fusion pore nucleation.
- Fusion pore expansion.

Our results show that the synaptotagmin-1 C2B Domain plays no significant role during the membrane fusion and fusion pore nucleation. However, C2B stabilizes the fusion pore once it is expanded, keeping it open without the need to apply any external force. The expanded fusion pore remains open under unbias conditions for 1.1 microseconds if the synaptotagmine is not present, while it remains stable for at least 10 microseconds for membranes containing one synaptotagmine.
This unexpected difference of 1 order of magnitude is clarified by describing protein–lipid interactions at the molecular level between different groups of protein and lipid molecules. After running a radial distribution function to quantify polybasic region coordination with the three species of lipids in the bilayers (1-palmitoyl2-oleoyl-glycero-3-phosphocholine: POPC, 1-palmitoyl-2-oleoyl-sn-glycero-3-phospho-L-serine: POPS and phosphatidylinositol4,5-bisphosphate:POP2) as a function of the distance we found that POP2 lipids were the ones that highly coordinate with the protein.

Conclusions:
We have demonstrated that the synaptotagmin-1 C2B domain has negligible effects on the membrane fusion and in the nucleation of the fusion pore. However, by expanding the fusion pore, we have shown that the C2B polybasic lysine-rich region strongly interacts with highly anionic phosphatidylinositol bisphosphate lipid molecules, driving the formation of POP2 aggregates and extending the life of the fusion pore.

Oral Presentation

Signal processing on graphs to measure similarity between gene annotations in the Gene Ontology

Tiago López, Leandro E. Di Persia, Diego H. Milone

sinc(i), Instituto de Investigación en Señales, Sistemas e Inteligencia Computacional, FICH/UNL-CONICET

Background:

The semantic similarity measures based on ontologies are useful in many applications, such as the inference of functions of genes or proteins annotated in the Gene Ontology. This application could improve the process of gene annotation, significantly reducing the number of laboratory experiments required. The classical measures for this application are based on the frequency of annotation of terms, but lack certain basic properties when considered as distances. As this could reduce the performance of an inference algorithm we propose to define new measures that incorporate more directly the structure of the Gene Ontology graph, applying the theory of signal processing on graphs.

Results:

Genes are represented as signals in a graph, defining paths that travel from the root node to the annotated terms. The proposed measures consist of defining dictionaries to transform the gene annotations and get the euclidean distances in this projected space. The dictionaries contain all the paths from the root node to the leaves, all the paths to each leaf combined or the eigenvectors of the graph Laplacian, that is, the Fourier Graph Transform atoms. These measures are evaluated by comparing the distances between genes annotated in a GO sub-ontology, and by the performance in the prediction of gene functions with a Bayesian approach. The distance for genes in a sub-ontology gave the expected results for the proposed measures. In the prediction task, the F1 score was higher for the proposed measures than the classical measures. Particularly, the measure based on the Graph Fourier Transform gave the higher performance scores for the automatic function prediction.

Conclusions:

The results show that the proposed measures adjust better to the notions of semantic similarity between genes, and are consistent with the mathematical properties of a distance. In the inference of gene functions, the proposed measures proved to be an appropriate alternative, even improving the performance of the classical measures.

Oral Presentation

ChronoRoot: High-throughput phenotyping by deep segmentation networks reveals novel temporal parameters of plant root system architecture

Nicolás Gaggion*, Federico Ariel**, Vladimir Daric***, Éric Lambert***, Simon Legendre***, Thomas Roulé***, Alejandra Camoirano**, Diego H Milone*, Martin Crespi***, Thomas Blein***, Enzo Ferrante*

* Research Institute for Signals, Systems and Artificial Intelligence, sinc(i), CONICET-UNL
** Instituto de Agrobiotecnología del Litoral, IAL, CONICET-UNL
*** Institute of Plant Sciences Paris-Saclay (IPS2), CRNS, INRIA

Background:

Deep learning methods have outperformed previous techniques in most computer vision tasks, including image-based plant phenotyping. However, massive data collection of root traits and the development of associated artificial intelligence approaches have been hampered by the inaccessibility of the rhizosphere. Here we present ChronoRoot, a system that combines 3D-printed open-hardware with deep segmentation networks for high temporal resolution phenotyping of plant roots in agarized medium.

Results:

We developed a novel deep learning–based root extraction method that leverages the latest advances in convolutional neural networks for image segmentation and incorporates temporal consistency into the root system architecture reconstruction process. Automatic extraction of phenotypic parameters from sequences of images allowed a comprehensive characterization of the root system growth dynamics. Furthermore, novel time-associated parameters emerged from the analysis of spectral features derived from temporal signals.

Conclusions:

We developed a novel deep learning–based root extraction method that leverages the latest advances in convolutional neural networks for image segmentation and incorporates temporal consistency into the root system architecture reconstruction process. Automatic extraction of phenotypic parameters from sequences of images allowed a comprehensive characterization of the root system growth dynamics. Furthermore, novel time-associated parameters emerged from the analysis of spectral features derived from temporal signals.

Poster Session: http://bit.ly/cab2c-2021-posters
Poster Gaggion et al: https://youtu.be/eGf113JnqHU

New insights in nematode comparative genomics

Lucas Luciano Maldonado, Pablo Gonzalez, Nahili Giorello, Mark Blaxter, Gisela Franchini and Laura Kamenetzky
IMPAM-UBA-CONICET
lkamenetzky@gmail.com

Dioctophyme renale known as “giant kidney worm”, is the biggest parasitic nematode, adult females can reach up to 103 cm long affecting several species of land mammals including humans. We generated high-coverage assembly combining short and long DNA reads and transcriptomes of dissected male and female parasites. We assemble 234 Mb in 219 contigs with 46.8% GC and ~58% of repetitive regions. Benchmarking of universal single copy orthologs estimates a completeness of 92.9% using eukaryota lineage dataset and 64.1% using nematoda lineage dataset. Similar values are found in related species such as Trichinella spiralis and Trichuris muris suggesting the presence of Clade I specific protein repertoire. More than 20,000 protein coding genes were functionally annotated. This genome information is highly informative for Clade I nematode which are still underrepresented in genome databases and will be placed in an evolutionary context by comparative genomics analysis.

Poster Session: http://bit.ly/cab2c-2021-posters
Poster Maldonado et al: https://youtu.be/5OldzcmbLqo

Automatic GO prediction of proteins on SARS-CoV-2

Elizabeth Chiacchiera[1], Elizabeth Tapia[1], Flavio E. Spetale[1]
1- CIFASIS-UNR/CONICET. Centro Internacional Franco-Argentino de Ciencias de la Información y de Sistemas.
spetale@cifasis-conicet.gov.ar

Background:

Gene Ontology (GO) is a structured repository of concepts (GO-terms) including three sub-ontologies, biological process (BP), molecular function (MF), and cellular component (CC). Although gene products should be ideally annotated simultaneously over the three sub-ontologies, with only a few exceptions, in-silico annotation methods work on individual sub-ontologies. Among the exceptions, annotation methods based on cross-ontology association rules and interaction networks can be mentioned. However, the applicability of these methods is somewhat limited, since association rules can only be used with GO transitive relationships and interaction networks need huge amounts of curated data only available for model organisms.

Results:

In this work, we predict the GO functionality of SARS-CoV-2 proteins using a novel graph-based Machine Learning package designed for the automatic annotation of protein coding genes across the three GO subdomains. The package, called FGGA (https://bioconductor.org/packages/fgga), provides fully interpretable graphical annotations amenable to expert analysis. A set of 8574 SARS-CoV-2 protein sequences was collated from the UniProt database based on their GO-terms and evidence codes. The FGGA annotation algorithm assembles individual GO term predictions issued by binary SVM classifiers. Regarding the training of individual SVMs, a minimum of 50 positively annotated protein sequences was considered. In addition, to assemble conveniently balanced training datasets, positively annotated protein sequences were complemented with negative annotated protein counterparts using an inclusive separation policy. Concerning characterization methods of individual protein sequences in terms of a fixed number of input features, the measurement of 478 physicochemical/secondary structure properties and sorting signals were considered. SARS-CoV-2 FGGA predictions were evaluated using a 5-fold cross-validation approach and hierarchical Precision, Recall and F-score performance metrics. For the set of 349 GO-terms (BP-CC-MF) analyzed, we obtained 90%, 93% and 91% of the hierarchical Precision, Recall and F-score metrics, respectively.

Conclusions:

The GO annotation of protein coding genes in viruses by well-established sequence similarity methods is a challenging task due to their fast-mutating nature. FGGA overcomes this critical limitation using the power of supervised Machine Learning methods. Preliminary results on SARS-CoV-2 proteins confirm the feasibility of our proposal.

Poster Session: http://bit.ly/cab2c-2021-posters
Poster Chiacchiera et al: https://youtu.be/aDAC9C-qTrs

Transcriptional alterations induced by INGAP-PP in rat islets

Macarena Algañaras[1], Ana C. Heidenreich [2], Agustín Romero [2], Carolina L. Román[1], Juan J. Gagliardino[1], Bárbara Maiztegui[1], Luis E. Flores[1], Santiago A. Rodríguez-Seguí [2,3*]

[1]CENEXA-Centro de Endocrinología Experimental y Aplicada (UNLP-CONICET LA PLATA; Centro Colaborador OPS/OMS para Diabetes), Facultad de Ciencias Médicas UNLP, La Plata, Argentina, [2]CONICET-Universidad de Buenos Aires, Instituto de Fisiología, Biología Molecular y Neurociencias (IFIBYNE), Buenos Aires, Argentina , [3]Universidad de Buenos Aires, Facultad de Ciencias Exactas y Naturales, Departamento de Fisiología y Biología Molecular y Celular, Buenos Aires, Argentina .

macarenaalga.aras@gmail.com, heidenreich.ac@gmail.com

Background:

Introduction: Type 2 Diabetes (T2D) is characterized by an impaired pancreatic β-cell mass/function together with an insulin resistant state. INGAP is a pancreatic protein that modulates β-cell mass and insulin secretion in presence of secretagogues, and its effects are also achieved by its derivate pentadecapeptide (INGAP-PP).

Aim: Uncover gene expression changes induced by the INGAP-PP rat islet treatment, and relate them to its known physiological effects.

Methods: RNA-seq data was generated from isolated rat pancreatic islets previously cultured with/ without INGAP-PP (n=3). Sequence read alignment was performed using HiSAT, followed by Stringtie for *de novo* gene annotation, transcript expression quantification and normalization. Public experimental data from rat brain, liver and islets were used to evaluate the tissue-specificity pattern of expression.

Results: A Linear Mixed Effects Model was applied to account for inter-replicate gene expression changes. This allowed identifying 2,008 genes whose expression was significantly modulated by INGAP-PP ($p<0.05$). Functional annotation of these differentially expressed genes was enriched in categories associated with neogenesis, apoptosis and insulin secretion signalling pathways. Expression of some non-annotated genes specifically expressed in rat β-cells was also modulated by INGAP-PP ($p<0.05$). Coding potential and similarity with other sequences annotated in public databases (BLAST) was performed to identify their potential function.

Conclusions:
INGAP-PP modulates gene expression of several mediators of signalling pathways related to neogenesis, β-cell activation, apoptosis, insulin secretion, among others that could modulate β-cell function/mass. The finding of non-annotated transcripts modulated by INGAP-PP presents a new challenge that we will further investigate to identify new targets that could be mediating the physiological effects induced by INGAP on rat islets.

Poster Session: http://bit.ly/cab2c-2021-posters
Poster Algañaras et al: https://youtu.be/mEbFV7J2QEE

Assembling of virulence-associated metabolic networks from *Pseudomonas aeruginosa* using genomic and proteomic data

González-Molero, W.*[a], Quintero-Troconis, E.[b], Acosta, H.[b], Rojas, A.[a]

[a]Centro Nacional de Cálculo Científico (CeCalCULA), Universidad de Los Andes, Mérida 5101, Venezuela.

[b]Laboratorio de Enzimología de Parásitos. Facultad de Ciencias, Universidad de Los Andes, Mérida 5101, Venezuela.

*fabiolamg123@gmail.com

Abstract

Pseudomonas aeruginosa is a gram-negative bacterium, characterized by its metabolic versatility and genetic plasticity, which allows it to reproduce in a wide range of environments, such as soils, water and food. It can be pathogenic or symbiont of different species of animals and plants. In recent years, the economic and biotechnological interest for this species has increased, due to its importance in medicine (virulence factors, resistance to antibiotics, toxin production) and also by it potential in other areas, such as, in agriculture, (producer of biofertilizers and pathogen control), and in bioremediation. In order to elucidate the metabolic potential of this species, an assembly of metabolic networks based on genomic and proteomic data was performed and compared with the respective metabolic pathways of three other species of the *Pseudomonas* genus: *P. putida*, *P. syringae* y *P. fluorescens,* these represent model organisms, with different lifestyles: phytopathogens, rhizobacteria and free-living, respectively. Based on metabolic databases (genes, proteins and metabolites) and bibliographic information, the metabolic pathways involved in the synthesis of siderophores, phenazines, quorum sensing and biofilm were evaluated and the metabolic networks were assembled. *P. aeruginosa*, has 3 unique compounds involved in the establishment and pathogenicity of the species, the siderophore pyocellin and the phenazine Pyocyanine (PYO) and 5-methyl-phenazine-1-carboxylate (5MPCA). In addition, it possesses a distinctive Quorum sensing signaling system from the other species, dependent on quinolones: 2-heptyl-4-quinolone (HHQ) and 2-heptyl-3-hydroxy-4-quinolone (PQS), which allows them cell communication and biofilm formation. The results confirm that these routes can be used as potential therapeutic targets to treat infections produced by this pathogenic bacterium.

PED: more findability and improved reusability of protein ensembles

Hatos A[1], Quaglia F[1,2], Piovesan D[1], Tosatto SCE[1]

[1]Department. of Biomedical Sciences, University of Padua, Padova 35131, Italy
[2]Institute of Biomembranes, Bioenergetics and Molecular Biotechnologies, CNR-IBIOM, Bari, Italy

Background:

The Protein Ensemble Database (PED, https://proteinensemble.org/) is an open access archive for the deposition of conformational ensembles, including intrinsically disordered proteins (IDPs). High quality of the data is guaranteed by a new submission process, which combines both automatic and manual evaluation steps. A team of biocurators integrate structured metadata describing the ensemble generation methodology, experimental constraints and conditions. Structural ensembles measured with NMR, SAXS, or FRET provide a broader representation of state of the art ensemble generation methods than the previous versions.

The new version, PED 4.0, is hosted at the University of Padua, Italy and has been completely redesigned and reimplemented with cutting-edge technology, thanks to a collaboration between European and Argentinian universities and research institutes part of the IDPfun consortium. Together with ELIXIR - the European bioinformatics infrastructure for life sciences - we provide more visibility to novel protein dynamics studies. Schema.org/BioSchemas mark-up allows search engines, e.g. Google dataset search, to find the deposited ensembles in PED. Recently PED joined the 3D-Beacons network - led by PDB in Europe - bringing together experimentally determined and predicted protein structure models from several providers, e.g. PDBe, PED, SASBDB, SWISS-MODEL and AlphaFold DB.

Results:

In this work, we present significant updates and upgrades to the new version of the PED, which now holds about eight times more entries compared to its first publication in 2014. By using the PED graphical interface, users can access descriptors of the qualitative and quantitative properties of the ensembles. We provide a new search engine that allows to build advanced queries and search the entry fields, e.g. cross-references to IDP-related resources (DisProt, MobiDB, BMRB and SASBDB), along with broader findability opportunities thanks to the implementation in BioSchemas and 3D-Beacons.

Conclusions:

We expect the renewed PED to play an even more crucial role for researchers interested in the atomic-level understanding of IDP function, thanks to the integration in the 3D-Beacons network, and be a central resource for the deposition of structural ensembles of IDPs.

Oral Presentation

Assessing Conservation of Alternative Splicing with Evolutionary Splicing Graphs

Diego Javier Zea[1], Hugues Richard[1,2] & Elodie Laine[1]

1 LCQB, Sorbonne Université, Paris, France
2 Robert Koch Institute, Berlin, Germany

Background:
Alternative splicing (AS) can significantly expand the proteome in eukaryotes by producing several transcript isoforms from the same gene. It has been associated with multiple biological functions and its deregulation with neurodegenerative disorders and cancer. Little is known about the contribution of alternative splicing to protein evolution. Furthermore, current analyses of sequence conservation fail to account for transcript variability across species. Our work introduces new concepts for the study of protein evolution, enabling for the first time granular estimates of alternative splicing conservation and significantly improving our knowledge of the amount of functionally relevant variations. We have developed ThorAxe (doi.org/gwxg), PhyloSofS (doi.org/gwxh), and Ases to help researchers understand the interplay between alternative splicing and protein evolution.

Results:
ThorAxe infers orthology relationships between exonic regions in the context of alternative splicing. We can use these orthologous exon groups to create an evolutionary splicing graph (ESG). The ESG summarises the transcript variability observed across species, allowing the detection of conserved alternative splicing events. We use them to show a clear link between the functional relevance, tissue-regulation, and conservation of alternative splicing events on a set of 50 human genes. Furthermore, ESGs constructed for the whole human protein-coding genome have allowed us to identify a few thousand genes where alternative splicing modulates the number and composition of pseudo-repeats shared across species. Then, PhyloSofS takes ThorAxe's outputs to infer a phylogenetic forest explaining the evolutionary relation between transcripts across species. Finally, the webserver Ases (www.lcqb.upmc.fr/Ases) integrates both tools facilitating the study of alternative splicing evolution.

Conclusions:
Ases, allow easy access to ThorAxe and PhyloSofS, two novel tools allowing for the analysis of protein evolution at the transcript level. ThorAxe is the central tool on that triad and introduces new concepts and representations that help us identify conserved alternative splicing events. We have applied that method on the whole human proteome, shedding light on possible functional events related to exonic duplication.

Oral Presentation

Target identification for repurposed drugs active against SARS-CoV-2 via inverse docking

Sergio R. Ribone[#], S. Alexis Paz[#], Cameron F. Abrams[^], and Marcos A.Villarreal[#*].

#Facultad de Ciencias Químicas, Universidad Nacional de Córdoba. Argentina. Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET). ^Department of Chemical and Biological Engineering, Drexel University, Philadelphia, United States.
*mvillarreal@unc.edu.ar

Background:
Screening already approved drugs for activity against a novel pathogen can be an important part of global rapid-response strategies in pandemics. Such high-throughput repurposing screens have already identified several existing drugs with potential to combat SARS-CoV-2. However, moving these hits forward for possible development into drugs specifically against this pathogen requires unambiguous identification of their corresponding targets, something the high-throughput screens are not typically designed to reveal.

Results:
We present here a new computational inverse-docking protocol that uses all-atom protein structures and a combination of docking methods to rank-order targets for each of several existing drugs for which a plurality of recent high-throughput screens detected anti-SARS-CoV-2 activity. We demonstrate validation of this method with known drug-target pairs, including both non-antiviral and antiviral compounds. We subjected 152 distinct drugs potentially suitable for repurposing to the inverse docking procedure. The most common preferential targets were the human enzymes TMPRSS2 and PIKfyve, followed by the viral enzymes Helicase and Plpro. All compounds that selected TMPRSS2 are known serine protease inhibitors, and those that selected PIKfyve are known tyrosine kinase inhibitors.

Conclusions:
Detailed structural analysis of the docking poses revealed important insights into why drug-target selections arose, and could potentially lead to more rational design of new drugs against these targets.

Oral Presentation

DisProt in 2022: improved quality and accessibility of protein intrinsic disorder annotation

Federica Quaglia[1,2], András Hatos[2], Edoardo Salladini[2], Damiano Piovesan[2], Silvio C.E. Tosatto[2]

[1] Institute of Biomembranes, Bioenergetics and Molecular Biotechnologies, National Research Council (CNR-IBIOM), Bari, Italy

[2] Department of Biomedical Sciences, University of Padova, Padova, Italy

Background:

DisProt, the Database of Intrinsically Disordered Proteins (https://disprot.org) is the major repository of manually curated annotations of intrinsically disordered proteins and regions from the literature. To reflect the steady progress of the protein disorder field, it is important to update and upgrade DisProt as the primary resources of manually curated, experimentally confirmed protein disorder. The previous release of the database, DisProt 8 contained about 1,500 entries and 3,500 disordered protein regions. In the current release, DisProt 9, not only did we increase these numbers, but also improved the reliability and quality of entries by introducing a reviewing process. Dissemination activities of the database content are advertised in the DisProt blog (https://disprot.org/blog/) and on our Twitter account (https://twitter.com/disprot_db). DisProt 9 presents a new graphical interface and updated features, such as the integration of two ontologies and the connection with APICURON (https://apicuron.org/), a database to credit and acknowledge the work of biocurators. With these improvements, DisProt continues to be a primary resource of protein disorder for the structural-molecular biology community.

Results:

We report here the updates in DisProt 9, including a restyled web interface, refactored Intrinsically Disordered Proteins Ontology (IDPO), improvements in the curation process and significant content growth of around 30%. Higher quality and consistency of annotations is provided by a newly implemented reviewing process and training of curators, with detailed curation guidelines and virtual training sessions. The increased curation capacity is fostered by the integration of DisProt with APICURON, a dedicated resource for the proper attribution and recognition of biocuration efforts. The DisProt technological infrastructure has been renewed to improve reproducibility by implementing exhaustive versioning of all entries. Better interoperability is provided through the adoption of the Minimum Information About Disorder (MIADE) standard and an active collaboration with the Gene Ontology (GO) and Evidence and Conclusion Ontology (ECO) consortia and the support of the ELIXIR infrastructure.

Conclusions:

DisProt serves the scientific community as a gold standard resource for IDP/IDR annotations. Compared to the previous version, it has improved data accessibility and quality, and significantly increased annotation volume. DisProt is well connected to other databases and consortia, and is active in the development of new standards and ontologies, integrating now GO and ECO. The long-term maintenance of DisProt is guaranteed by its central role within the European Union’s Horizon 2020 IDPfun program and the ELIXIR IDP Community, the reference scientific communities involved in the study of intrinsically disordered proteins.

Oral Presentation

The development of a highly multiplex long-read sequencing protocol for SARS-CoV-2 genomes based on NS-watermark barcodes

García Labari I.[1], Ezpeleta J.[1,2], Posner V.[3,4], Villanova G.V.[3,4], Lavista-Llanos S.[1], Bulacio P.[1,2], Spetale F.[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], Casal P.E.[4], Spinelli S.[7], Giri A.A.[4,6], Arranz S.[3], Tapia E.[1,2]

1- CIFASIS-UNR/CONICET. Centro Internacional Franco-Argentino de Ciencias de la Información y de Sistemas.
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.
garcialabari@cifasis-conicet.gov.ar

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 protocol for whole-genome SARS-CoV-2 sequencing with compatible costs, accessibility, and processing times to the demands of the COVID-19 emergency.

Results:

We build upon the amplicon tiling strategy described previously by Quick J. et al 2017 for the rapid whole-genome virus sequencing of clinical samples. We looked for an alternative to the established SARS-CoV-2 ARTIC (amplicon-tiling) sequencing protocol that could take full advantage of portable sequencing machines of low-capital access cost now available in the emerging market of long-read sequencing technologies. 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 for the parallel sequencing of thousands of genomes instead of the dozens currently reported in the literature. Based on the NS-watermark barcoding approach described previously by Ezpeleta J. et al 2017, we selected 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 were sequenced together on the MinION sequencer. After base-calling and trimming of sequencing adapters, reads were individually deconvoluted with an *in-house* script prepared for calling the NS-watermark decoding software. Even for 96 samples, high coverage rates (> 98% of the genome) and depths (> 30X in each amplicon fragment of 1.5 Kb) were obtained. We uploaded more than 200 genome high-quality sequences to the GISAID database, including the first complete SARS-CoV-2 genome from Santa Fe.

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. **Funding:** Covid Federal (SF-11), Santa Fe-Argentina, and Argentag SAS - www.argentag.com

Oral Presentation

Prediction of gene silencing in *Arabidopsis thaliana* using decision trees and support vectors machines algorithms

Pozzi F.I.[1,2], Felitti S.A.[2]

[1] Cátedra de Microbiología. Facultad de Ciencias Agrarias, Universidad Nacional de Rosario, S2125ZAA Zavalla, Santa Fe, Argentina. [2] Instituto de Investigaciones en Ciencias Agrarias de Rosario. Facultad de Ciencias Agrarias (UNR). Campo Experimental Villarino, CC14, (S2123ZAA), Zavalla, Santa Fe, Argentina. pozzi@iicar-conicet.gob.ar

Background:

Although most angiosperms require an endospermic balance number (EBN) for normal endosperm development, tetraploid *Paspalum notatum* is EBN- insensitive. A candidate gene (GG13) of *Paspalum notatum* associated with endosperm development in insensitive EBN crosses was analyzed by gene silencing in *Arabidopsis thaliana*. When comparing the silenced (S) and control (C) conditions, S condition evidenced less relative expression in RT-qPCR experiments and a more elongated shape (shape index 1.83 vs. 1.57) in phenotypic analyses. Length (L), width (W) and shape index (L / W) variables were evaluated. The objective of the work was 1- To predict, from 1896 phenotypic data, the classes: S and C, through the use of the Decision Trees (DT) and Support Vector Machines (SVM) algorithms. 2- Determine which variable has more weight in class separation (S and C), in order to optimize the number of variables to be evaluated in future experiments for the GG13 gene, optimizing work time and effort. Predictive models were evaluated with RStudio

Results:

In this work it was obtained that in the DT model the variable with the greatest weight to achieve the separation of classes S and C was L/W (50), while W and L would have the same importance or weight to predict (25). In addition to the above, it was possible to define that 63% of the data is above the L/W value: 1,595 (value defined in primary splits). For this model the precision was 79% and the error 21%. The optimal value of the cp parameter was 0.01. For the SVM model, the total of support vectors was 633. For this model the precision was 72% and the error 28%. The optimal value of the cost parameter was 0.001.

Conclusions:

The results of the analysis of the predictive models DT and SVM we can say that: precision of both between scarce and good in terms of their power of generalization. The predictive ability of class separation and generalizability of the DT model outperformed SVM. It was not possible to optimize the number of variables to be evaluated in future experiments for the gene under study, because the variable that best separates it is L/W, which to be determined requires the values of variables: L and W (variables that showed the same weight).

Poster Session: http://bit.ly/cab2c-2021-posters
Poster Pozzi et al: https://youtu.be/AwvXfE2g6LI

What determines the evolutionary rate of amyloid proteins?

Diego Javier Zea, Guillermo Benítez, Juan Mac Donagh, Cristian Guisande Donadio, Nicolas Palopoli, Julia Marchetti, Maria Silvina Fornasari, Gustavo Parisi

Departamento de Ciencia y Tecnología, UNQ, CONICET, Buenos Aires, Argentina

Background:

Amyloid fibrils are insoluble, closely packed, and highly ordered arrangements that most proteins can adopt, although with different tendencies. They arise when a soluble protein or protein fragment aggregates as a filament. Filament's monomers stick between them through an array of β-sheets perpendicular to the fiber's axis. There are plenty of studies associating amyloids with human diseases. However, amyloids can also have functional roles in humans thanks to their high stability of the amyloid state. Sadly, few studies are focussing on the evolution of biologically functional and pathological human amyloids.

Results:

This work studies the evolutionary rates of 65 human proteins with reported capacity to form functional or pathological amyloids compared to others. Despite being highly expressed, we have found that functional and pathological amyloids are also fast-evolving proteins. Therefore, amyloids behave differently from most proteins in evolutionary terms. We performed a multivariate analysis, including many factors affecting the evolutionary rates of proteins. That analysis shows that the main determinants of amyloids' evolutionary rates are cellular location, the number of disulfide bonds, and the interaction with chaperones. We have also tested the disulfide bond and chaperone hypotheses using ancestral reconstruction. In particular, we have measured the ddG of the reconstructed mutations to see whether the amyloids behave like other human proteins sharing those characteristics.

Conclusions:

These results show that most amyloids are located outside the cytoplasm, avoiding the purifying pressure acting on other highly expressed proteins to prevent cytotoxicity. At the same time, their disulfide bonds and their interactions with intra- and extra-cellular chaperones stabilize their structures, allowing them to accumulate mutations. The previously mentioned factors promote the accumulation of destabilizing mutations on amyloids, further increasing their amyloidogenic propensity. Chaperones are particularly interesting, as they form a vicious circle; mutated amyloid proteins request extensive chaperone assistance to avoid aggregation, but their chaperone interaction could favor the accumulation of destabilizing mutations.

Oral Presentation

Impact of protein conformational diversity on AlphaFold predictions

Tadeo Saldaño*[1], Nahuel Escobedo*[1], Diego Javier Zea5, Juan Mac Donagh[1], Julia Marchetti[1], Ana Julia Velez Rueda[1], Eduardo Gonik[4], Agustina García Melani[3], Julieta Novomisky Nechcoff[1], Martin Salas[1], Tomás Peters[2], Nicolás Demitroff[2], Sebastian Fernandez Alberti[1], Nicolas Palopoli[1], Maria Silvina Fornasari[1] and Gustavo Parisi[1#].

*authors contributed equally
#corresponding author: Gustavo Parisi, gusparisi@gmail.com

1 Departamento de Ciencia y Tecnología, Universidad Nacional de Quilmes, Bernal, Argentina
2 Fundación Instituto Leloir-Instituto de Investigaciones Bioquímicas de Buenos Aires
3 IMBICE (CONICET - UNLP). Laboratorio de Electrofisiología
4 INIFTA (CONICET-UNLP) - Fotoquímica y Nanomateriales para el ambiente y la biología (nanoFOT)
5 Independent researcher, 14 rue Léo Lagrange, 31400, Toulouse, France

Abstract

After the outstanding breakthrough of AlphaFold in predicting protein 3D models, new questions appeared and still remain unanswered. The ensemble nature of proteins challenges the structural prediction problem because under this paradigm, predicted models should represent a set of conformers instead of single structures. Essentially, the evolutionary information captured by deep learning techniques should be able to resolve the degeneracy present in single sequences adopting several conformations. Here, we explore AlphaFold2 predictions under this ensemble paradigm. Using a collection of hand-curated apo-holo conformations, we found that AlphaFold2 mostly predicts the holo form of a protein (70% of the cases) being unable to reproduce the observed conformational diversity (only in 2% of the cases). More importantly, we also found that AlphaFold2 performance worsens with increasing conformational diversity of the protein being studied. This impairment is related with the heterogeneity of conformational diversity found in the different members of the homologous family of the protein under study. Finally, we found that highly flexible regions between apo-holo conformations show a negative correlation with the local quality score (plDDT) indicating that plDDT in a single 3D model could be used to infer local conformational changes between apo and holo forms.

Poster Session: http://bit.ly/cab2c-2021-posters
Poster Saldaño et al: https://youtu.be/IEvQnEVL_pU

Molecular docking to explain effect of novel disease-causing variants in the *GP1BA* gene related to Bernard Soulier syndrome

Débora M Primrose PhD[a], Adriana I Woods PhD[b], Juvenal Paiva MS[c], Analia Sánchez-Luceros MD, PhD[bc]

[a]Higher School of Engineering, Informatics and Agri-food Sciences, University of Morón. Buenos Aires, Argentina.

[b]Laboratory of Hemostasis and Thrombosis, IMEX-CONICET-National Academy of Medicine. Buenos Aires City, Argentina

[c]Department of Hemostasis and Thrombosis, Hematological Research Institute, National Academy of Medicine. Buenos Aires City, Argentina

Background: Bernard Soulier syndrome (BSS) is a rare autosomal inherited disorder caused by homozygous or compound heterozygous disease-causing variants (DCVs) in any of the genes encoding for glycoproteins (GP) Ib (GPIb) (*GP1BA, GP1BB*) and GPIX (*GP9*). We described two novel DCVs, p.Arg80Gly and p.Tyr231Cys in the *GP1BA* gene related tSS in two family members with macrothrombocytopenia.
Aims: To predict the effect of the GPIbα-p.Arg80Gly and GPIbα-p.Tyr231Cys on their interactions with the von Willebrand Factor-A1 domain (VWF-A1D) by homology modelling and protein-protein docking.
Methods: The models GPIbα-p.Arg80Gly and GPIbα-p.Tyr231Cys were made by replacing Arg80 with Gly, and Tyr231 with Cys in wild-type-GPIbα (Swiss-PDBviewer). Docking between VWF-A1D and GPIbα mutants was obtained (PatchDock) and compared to wild-type-GPIbα-VWF-A1D. Hydrogen-bonds and root mean square deviation (RMSD) between α-carbon chains were obtained (UCSF Chimera).
Results: In-silico model of wild-type-GPIbα-VWF-A1D shows differences with GPIbα- p.Arg80Cys-VWF-A1 and p.Tyr231Cys-VWF-A1. Total RMSD between models is 4.85Å for GPIbα-p.Arg80Gly (0.02Å between GPIbα portions; 4.85Å between VWF-A1D). There is no change in the intramolecular hydrogen-bond between Arg80 and His86 of GPIb. The GPIbα-p.Arg80Gly causes a slight change in the molecule, altering the way it joins the VWF-A1D. This residue does not form hydrogen-bonds with the VWF-A1D. It was predicted as damaging by Provean and SIFT4G, possibly damaging by PolyPhen-2, and uncertain significance by Varsome, but benign by other 22 in-silico methods. I-mutant predicted this change as a large decrease in protein stability. The change Tyr231Cys causes a slight alteration in the conformation of GPIbα respecting wild-type-GPIbα. The RMSD is 38.05Å (0.65Å between GPIbα portions; 15.07Å between VWF-A1D). This change does not affect neighboring residues but alters the structure of GPIbα at the site that acts as a hinge, increasing the distance of the beta-switch region of GPIbα (flexible loop at Val243-Ser257). This modifies the interaction with VWF-A1D, changing hydrogen-bonds between these molecules. It was predicted as damaging by PolyPhen-2, SIFT, SIFT4G, Mutation Taster, MutPred, Provean, and M-Cap. I-mutant predicted this change as a large decrease in protein stability.
Conclusion: In-silico predictions were useful to demonstrate the protein interaction modifications caused by the changes in the GP1BA gene that lead to p.Arg80Gly and p.Tyr231Cys. These DVCs modify the tertiary structure of GPIbα affecting either its interaction with VWF-A1D or preventing the conformational change of the beta-switch of GPIbα from compact to extended form thus altering the VWF-A1D-GPIbα binding.

Poster Session: http://bit.ly/cab2c-2021-posters
Poster Primrose et al: https://youtu.be/1fWv7jcRM3M

Intrinsically Disordered Protein Ensembles Shape Evolutionary Rates Revealing Conformational Patterns

Julia Marchetti[1], Nicolás Palopoli[1], Alexander M. Monzon[2], Diego J. Zea[3], Silvio C.E. Tosatto[2], Maria S. Fornasari[1] and Gustavo Parisi[1]

1 Departamento de Ciencia y Tecnología, Universidad Nacional de Quilmes, CONICET, Bernal, Buenos Aires, Argentina. 2 Department of Biomedical Sciences, University of Padua, Padua, Italy. 3 Laboratoire de Biologie Computationnelle et Quantitative (LCQB), Sorbonne Université, IBPS, CNRS, Paris, France

Background:

Intrinsically disordered proteins (IDPs) lack stable tertiary structure under physiological conditions. The unique composition and complex dynamical behaviour of IDPs make them a challenge for structural biology and molecular evolution studies. Here we propose an evolutionary approach to describe the large conformational diversity present in IDPs. The aim of this work is to understand the direct impact on evolutionary rates of structural disorder.

Results:

We used 311 different NMR ensembles to explore the structural constraints imposed on evolutionary rates by residue-specific contacts observed in each conformer from the ensemble. We found that IDPs evolve under a strong evolutionary rate heterogeneity mainly originated by differences in their inter-residue contacts.

We also explored if the structural environment of a position could have an influence in its evolutionary rate, the contacts established with other residues and its observed flexibility (measured by RMSF). We found that these properties change as a function of the distance to the nearest ordered residue, suggesting that the influence of structural constraints imposed by ordered positions also extends to disordered sites.

Then we studied the evolutionary rates in more structural specific contexts and different domain architectures. We collected the evolutionary rate profiles and the RMSF profiles of proteins with a particular disorder-order architecture and we found that both rate profiles and RMSF profiles correlate with the observed conformational variability of the protein, thus allowing us to define different structure-function relationships from the site-specific evolutionary rates. Interestingly, our results also showed that correlation between rates and contact information significatively improves when structural information is taken from the ensemble and not from any individual conformer.

Conclusions:

Our results show that structural properties within disordered regions constrain evolutionary rates to conserve the dynamic behaviour of the conformational ensemble, highlighting the importance of structural analysis in unstructured proteins.

Oral Presentation

DRUG REPURPOSING TO FIND INHIBITORS OF SARS-CoV-2 MAIN PROTEASE

Conti G, Gomez Chavez JL, Angelina EL, Peruchena NM

Lab. Estructura Molecular y Propiedades, IQUIBA-NEA, Universidad Nacional del Nordeste, CONICET, Av. Libertad 5470, Corrientes 3400, Argentina

Background:

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is the strain of coronavirus that causes coronavirus disease 2019 (COVID-19), the respiratory illness responsible for the COVID-19 pandemic. As of October 2021, with about 6 billion doses of COVID-19 vaccines administered globally, the world is slowly returning back to normality. However, there is still a need for novel therapeutics for people that contracted the disease either because they were not vaccinated or because the vaccine was not effective for them.

Drug repurposing is a strategy for identifying new uses for approved drugs that has the advantage, over conventional approaches that attempt to develop a drug from scratch, that the time frame of the overall process can be significantly reduced because of the few number of clinical trials required.

In this work, a structure-based virtual screening of FDA-approved drugs was performed for repositioning as potential inhibitors of the main protease Mpro of SARS-CoV-2.

Results:

12 drugs were prioritized from the Virtual Screening campaigns as potential inhibitors of the Mpro enzyme. Some of the selected compounds turned out to be antiviral drugs already being tested in COVID-19 clinical trials or used to alleviate symptoms of the disease. Curiously, the most promising candidate is the naturally occurring broad spectrum antibiotic Oxytetracycline (OTC). This drug has largely outperformed the remaining selected candidates along all filtering steps of the virtual screening workflow. The closely related tetracycline-derived drug Doxycycline (DOX) recently has proven to reduce the viral load in Vero cells infected with SARS-CoV-2. Since OTC but not DOX surpassed the more stringent filters in the late stages of our virtual screening pipeline, we suspect that OTC will show even higher antiviral activity than DOX in viral replication experiments.

Conclusions:

Considering our computational findings together with the proven antiviral properties of DOX, we believe it is worth testing OTC in prospective viral replication studies. We encourage the scientific community working on COVID-19 projects to include this repurposing candidate on their experimental screening pipelines.

Poster Session: http://bit.ly/cab2c-2021-posters
Poster Conti et al: https://youtu.be/WDp9o07Hjvc

PCA of the FtsZ/Tubulin superfamily reveals the relationship between physicochemical features and the domain composition.

María del Saz Navarro[1], Damien Paul Devos[1].

1. Centro Andaluz de Biología del Desarrollo (CABD)-CSIC, Pablo de Olavide University, Spain.

Corresponding author e-mail: mariadsn89@gmail.com

Abstract: One of the main elements of the cytoskeleton of eukaryotic cells is the microtubule-forming tubulin, whose prokaryotic counterpart is the FtsZ protein. In a previous study we focused on the phylogenetic relationship between the proteins that gave rise to eukaryotic tubulin, which encompasses all the proteins of the FtsZ/Tubulin superfamily. To carry out the phylogenetic analysis study we use tools based mainly on sequence homology. We have recently extracted a set of biophysical characteristics and attributes derived from the sequences of this superfamily with the aim of finding out whether these characteristics allow us to support the grouping of the families based on sequence homology. For this purpose, we have carried out a principal component analysis study and used different clustering algorithms. Our aim with the clustering analysis is to find out if these sequence-derived features contain enough information to group the different proteins according to the families they belong to by similarity; those proteins that belong to the same family have similar features. We conclude that the extracted characteristics allow to group the different families according to the distribution of domains protein even if this information was not used during the training. On the other hand, if more than one family has the same domain distribution, it does not establish clear differences between them.

Poster Session: http://bit.ly/cab2c-2021-posters
Poster Navarro et al: https://youtu.be/2nZWZCW-EIw

DisPhaseDB: an integrative database of phase separation proteins and disease associated mutations

Alvaro Navarro[1], Fernando Orti[1], Elizabeth Martinez-Perez[1], Cristina Marino-Buslje[1*] and Javier Iserte[1*]

[1] Bioinformátics Unitl, Fundación Instituto Leloir. Av. Patricias Argentinas 435, Buenos Aires, Argentina

[*] Contributed equally

Background:

Cells compartmentalize biological processes to achieve spatial and temporal control of biochemical reactions. This is accomplished through organelles. Membrane-less organelles (MLOs) have recently gained relevance given their multiple essential roles in cells. MLOs are condensates of proteins and nucleic acids formed through liquid-liquid phase separation (LLPS) within the cellular milieu, allowing the differential concentration of macromolecules. Recent advances have shown that phase separation underlies many biological processes.

There are several molecular drives of phase separation, for example protein-protein or protein-RNA interaction domains, protein intrinsically disordered or low complexity regions (IDRs and LcRs, respectively), weak transient interactions, electrostatic interactions, pi-pi and cation-pi, among others.

The formation, composition and dynamics of MLOs are fine tuned processes whose disturbance can cause physiological liquid condensates to turn into pathological aggregates leading to diseases. This is the case of various pathologies such as Alzheimer, Parkinson, Huntington, ALS, cancer and infectious diseases among others. Therefore, it is not surprising that mutations in proteins associated with MLO can alter the molecular mechanisms driving various disorders.

There are several databases dedicated to LLPS proteins and MLOS, as well as to human mutations and diseases. However there is not a centralized and easy to use tool to access data. In order to fill this gap, we develop a new database, called **DisPhaseDB**, that focuses on this issue.

Results:

We present a metadatabase merging human LLPS proteins from dedicated databases such as PhasePro, PhaSepDB and DrLLPS and their mutations from DisGeNET, OMIM, Cosmic and ClinVar. We analyzed the records, unified and completed their metadata. For each protein we added information from Uniprot, domain information from PFAM, Prion-Like predicted domains, disorder content and Lc regions from MobiDB 3.0 and post-translational modification from PhosphoSitePlus. All the features are mapped onto the protein sequence and displayed by a friendly, easy to interpret visualization system. In such a way allowing users to locate at a first glance different attributes that might be of interest. For example, mutations occurring in a disorder region or within a particular domain. The database allows users to customize the subset of proteins to be retrieved by querying the database with several filters: MLO location, disease, protein accession among others.

Conclusions:

Here we introduce **DisPhaseDB**, a database of disease mutations found in MLO related LLPS proteins. We merge 3 LLPS proteins, and 4 variant databases together with extensive metadata. All the information is mapped onto the protein sequence and displayed with an easy to interpret visualization. DisPhaseDB currently contains 5.332 human proteins, 1606.105 mutations and 4.051 diseases. We use standard modem tools to develop the frontend (Angular+node.js+Bootstrap) and the backend (Python +Flask+MySQL). DisPhaseDB is available online at http://disphasedb.leloir.org.ar/.

Oral Presentation

VIRTUAL SCREENING OF PLANTS EXTRACTS WITH ANTI-CHAGASIC ACTIVITY GUIDED BY OMICS DATA

Gomez Chavez JL, Conti G, Angelina EL, Peruchena NM

Lab. Estructura Molecular y Propiedades, IQUIBA-NEA, Universidad Nacional del Nordeste, CONICET, Av. Libertad 5470, Corrientes 3400, Argentina

Background:

Chagas disease is caused by the protozoan parasite *Trypanosoma cruzi* and their treatment consists in the use of two nitro heterocyclic compounds that both present high toxicity and low effectiveness during the chronic phase.

Cymbopogon citratus extracts have been shown to alleviate the chronic phase symptoms of the disease, reducing amastigote nests and inflammatory infiltrates in the heart tissue of infected mice.

In this work we used bioinformatics tools to mine public functional genomics databases along with chemoinformatics tools and molecular docking to understand the action mechanism at the molecular level of *C. citratus* extract in chagasic cardiomyopathy.

Results:

The GSE41089 dataset from NCBI GEO contains 14154 genes that were processed for analysis. Genes with log Fold Change > 1 and adjusted p-value < 0.05 were considered over-expressed, resulting in 1465 that met the conditions for further analysis.

Functional gene enrichment analysis prioritized Gene Ontology terms associated with innate immune response and the most significatives biological pathways affected were those relationed to release of cytokines.

Genes in the top-enriched pathways that are known to be affected by the *C. citratus* ingredients, were selected as potential protein targets, resulting in 6 candidates.

Molecular docking of active plant ingredients were performed on the candidate targets to identify potential inhibitors. Promising results were found particularly for one of the candidate targets, Ptgs2 (prostaglandin-endoperoxide synthase 2). Compounds such as Nerolidol, Farnesol and Oleic Acid show quite similar binding modes and comparable anchoring strength than that of the enzyme substrate.

Conclusions

The results show that the combination of omics analysis with molecular docking is a promising strategy to understand the action mechanisms of plants extracts against chagas disease. The information obtained will allow us to search for healthier and less harmful treatments against *T. cruzi*.

Poster Session: http://bit.ly/cab2c-2021-posters
Poster Gomez Chavez et al: https://youtu.be/fAG-H_rwy2g

Systematic identification and prioritization of SLiMs that interact with the pocket protein family using Phage Display.

Lorenze C[1]; Safranchik M[1]; Garrone N[1]; Glavina J[1]; Chemes LB[1]

[1]Protein Structure-Plasticity and Function Laboratory. Instituto de Investigaciones Biotecnológicas (IIBIO-CONICET) (IIB-UNSAM). San Martín, Argentina

Background:
Short Linear Motifs (SLiMs) are short modular elements (3 to 10 amino acids), most commonly found in intrinsically disordered regions of proteins, that mediate protein-protein interactions. The pocket protein family includes the retinoblastoma (Rb), p107 and p130 proteins. The pocket proteins have a highly conserved "pocket" domain that interacts with their targets through the E2F and LxCxE SLiMs. In this study, we use bioinformatics tools to study the enrichment of these two SLiMs in peptides detected by a Phage Display assay using the pocket proteins as bait.

Results:
We used the Phage Display technique and the pocket domains from the pocket proteins to identify unknown targets. The phage library contains over a million peptides of 16 residues each, from disordered regions of the human proteome. Screening using the Rb, p107 and p130 pocket domains revealed 267, 95 and 513 hits respectively. As a first approach, we used two SLiM detection algorithms, SLiMFinder and MEME, that identify overrepresented motifs in a set of sequences.
SLiMFinder and MEME, identified an enrichment of the E2F motif (25% and 10% respectively) and the LxCxE motif (7% and 17% respectively) in the Rb screening. However, in the case of the screening with p107, we observed an enrichment only for the LxCxE motif (45% and 47%, respectively). No enrichment for known SLiMs was observed for the p130 protein, probably because of a poor p130 purification. These results suggest that it is possible to identify new proteins carrying functional motifs that interact with pocket proteins. Our next steps are the identification and scoring of LxCxE and E2F motifs by analyzing regular expressions, structure flexibility, PFAM domains, and punctuation systems such as FoldX and PSSM matrices in order to prioritize interactors for future *in vitro* validations.

Conclusions:
We were able to conduct an initial screening of pocket protein binding SLiMs and plan to extend the analysis using methods that allow the identification of peptides with a higher likelihood of being functional interactors. This will help expand the known interactome of the pocket protein family and the knowledge of its functions.

Poster Session: http://bit.ly/cab2c-2021-posters
Poster Lorenze et al: https://youtu.be/8K81ES4buEc

Predicting GO annotations by integrating protein knowledge with end-to-end deep learning

Gabriela A. Merino [1,2], Rabie Saidi[3], Diego H. Milone [2], María J. Martin[3], Georgina Stegmayer [2].

[1]Instituto de Investigación y Desarrollo en Bioingeniería y Bioinformática (IBB)-CONICET-UNER, Oro Verde, Argentina.
[2]Instituto de Investigación en Señales, Sistemas e Inteligencia Computacional (sinc(i))-CONICET-UNL, Santa Fe, Argentina.
[3]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Hinxton, United Kingdom.

Background:
Experimental validation and manual curation are the most precise ways for assigning Gene Ontology (GO) terms describing protein functions, but they are expensive, time-consuming, and cannot cope with the exponential growth of available data. Computational models for automatic function prediction are being constantly developed. However, their performance is still subject to improvement, especially for no-knowledge (NK) proteins, which have not been previously annotated.

Results:
We propose a novel end-to-end deep learning model for predicting GO terms by integrating different types of protein data. Our model is based on a feed-forward deep neural network involving several encoding sub-networks that learn specific features from each protein data, and a classification sub-network aimed to predict the full set GO terms. Here, we used multiple features from sequence and organisms taxa as input, and trained and evaluated our model for predicting annotations of NK proteins following the CAFA challenge setup. As a data augmentation strategy aimed to reduce the imbalance differences between train and evaluation sets, NK proteins were augmented using training proteins with no changes in their annotations up to the challenge deadline. CAFA3 benchmark proteins were used for evaluating our model, obtaining F-max scores of 0.34, 0.55, and 0.55 for biological process (BP), cellular component (CC), and molecular function (MF) sub-ontologies, respectively. These results revealed our model performed in the top 5 CAFA3 methods, achieving very competitive scores with respect to those of the best competitors for BP and CC. It was also the second-best method when predicting MF. Even more, our model achieved scores higher than those reported by the deep learning methods of the state-of-the-art, for NK proteins.

Conclusions:
The obtained results proved that end-to-end deep learning models are able to reliably predict likely annotations for proteins, with high precision and without any restriction in the number of GO terms to predict, enhancing the discovery of new functions. It has been found that integrating protein knowledge and using data augmentation during model training are effective approaches for improving the prediction of GO terms describing protein functions.

Poster Session: http://bit.ly/cab2c-2021-posters
Poster Merino et al: https://youtu.be/CY4_OG3N6Cc

Complex networks integration for uncovering alternative splicing temperature related regulatory patterns in Arabidopsis thaliana

Andres Rabinovich[1] and Ariel Chernomoretz[2]
[1]Fundación Instituto Leloir
[2]University of Buenos Aires
andresrabinovich@gmail.com

BACKGROUND: Alternative splicing (AS) has been proposed as a post-transcriptional regulatory mechanism to increase transcript diversity in eukaryotic organisms. AS involves the interaction of splicing factors with their targets in a coordinated and combinatorial way in a fashion that is currently unclear.

RESULTS: In this work we studied organizational patterns that emerge on the co-splicing level and its relationship with transcriptional regulation in order to understand subgenic regulation at the systems level. We consider an Arabidopsis thaliana temperature RNA-Seq dataset consisting of a time course for 2 different temperatures, 22º and 4º, measured along 2 days, every 3 hours. Using ASpli, an RNA-Seq analysis pipeline developed in our lab, we first characterized changes in splicing patterns between 22º and 4º. Then, using a Random Forest based scheme we built a multilayer network consisting of a transcriptional layer, a co-splicing layer and the relationship between them, for each temperature, and compared both networks to uncover changes in temperature related splicing regulatory patterns.

CONCLUSIONS: Using ASpli, an RNA-Seq analysis pipeline developed in our lab, we were able to build and integrate co-splicing and transcription factors networks, uncovering temperature related regulatory patterns in A. thaliana with possible biological relevance

Poster Session: http://bit.ly/cab2c-2021-posters
Poster Rabinovich et al: https://youtu.be/uo1fEieN_0k

Modeling HNF1B-associated monogenic diabetes using human iPSCs reveals an early stage impairment of the pancreatic developmental program

Evelyn Olszanowski[1,2,Φ], Ranna El-Khairi[3,4,5,Φ], Daniele Muraro[3,4], Pedro Madrigal[3,4], Ludovic Vallier,[3,4,5,θ] and Santiago A. Rodríguez-Seguí[1,6,θ]

[1]Instituto de Fisiología, Biología Molecular y Neurociencias (IFIBYNE), CONICET-Universidad de Buenos Aires, Ciudad Universitaria, Buenos Aires, Argentina

[2]Departamento de Química Biológica, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Buenos Aires, Argentina

[3]Wellcome Medical Research Council Cambridge Stem Cell Institute, Anne McLaren Laboratory for Regenerative Medicine, University of Cambridge, Cambridge, UK

[4]Wellcome Sanger Institute, Hinxton, Cambridge, UK

[5]Department of Surgery, University of Cambridge, Cambridge, UK

[6]Departamento de Fisiología, Biología Molecular y Celular, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Buenos Aires, Argentina

Φ Co-first authors

θ Co-corresponding authors: lv225@cam.ac.uk, srodriguez@fbmc.fcen.uba.ar

Background:

Maturity-onset diabetes of the young (MODY) is the most common form of monogenic diabetes, and it is characterized by autosomal dominant inheritance, and hyperglycemia due to β cell failure. Hepatic nuclear factor 1b (HNF1B), plays an important role in the normal development of the kidney, liver, pancreas, bile ducts, and urogenital tract, through tissue-specific regulation of gene expression in these organs. In humans, heterozygous mutations in HNF1B result in a multisystem disorder, associated with MODY5.

Results:

Bulk RNA-seq experiments of HNF1B$^{-/-}$ (1βHom) hiPSC-derived progenitors at the FP stage (day 6), and at posterior foregut (day 8) revealed upregulation in alternative developmental pathways, notably heart, kidney, and nervous system development, showing the central role of HNF1B in the specification of the foregut toward the pancreatic lineages. In all 1βHet and 1βHom samples from all stages we found a consistent downregulation of the HNF1A antisense long non-coding RNA (lncRNA HNF1A-AS1). Single-cell RNA-seq on HNF1B$^{+/+}$ (1βWT) and HNF1B$^{+/-}$ (1βHet) cells from day 13 revealed that the number of less proliferative (late) MPCs in D13-1βHet samples was higher than in their 1βWT counterparts, and this increase appeared to take place mainly at the expense of the highly proliferative (early) MPC population, which express increased levels of the proliferation markers TOP2A and AURKB. These results suggest that HNF1B plays an important role in allowing the proliferative early MPC stage.

Conclusions:

Our results show that lack of HNF1B blocks specification of pancreatic fate from the foregut progenitor (FP) stage, but HNF1B haploinsufficiency allows differentiation of multipotent pancreatic progenitor cells (MPCs) and insulin-secreting b-like cells. We show that HNF1B plays a central role in the specification of the foregut towards the pancreatic lineages by controlling key master regulators; the absence of HNF1B affects foregut patterning by allowing cells to adopt alternative fates. Also, we found that HNF1B haploinsufficiency impairs cell proliferation in FPs and MPCs. This could be attributed to impaired induction of key pancreatic developmental genes, including SOX11, ROBO2, and additional TEAD1 target genes whose function is associated with MPC self-renewal.

Poster Session: http://bit.ly/cab2c-2021-posters
Poster Olszanowski et al: https://youtu.be/YviCt0GYG1Y

Genotyping by sequencing for molecular characterization of the VArg1 and VArg2 *Verticillium dahliae* pathogenic strains of sunflower.

Aguilera, P.N. (1)*; Montecchia, J.F. (1); Ben Guerrero, E. (1); Quiroz F. (2); Heinz R. (1);
Filippi C. (1); Troglia C. (2); Lia, V. (1); Martínez, M.C. (1); Paniego, N. (1)

(1) Instituto de Agrobiotecnología y Biología Molecular, UEDD INTA-CONICET, Argentina.
(2) EEA Balcarce, INTA, Argentina

Background:

Sunflower leaf mottle and wilt caused by *Verticillium dahliae* Kleb. is one of the most important diseases of the crop. Genetic resistance is the most effective strategy for controlling this soil-borne ascomycete fungus. Two local phytopathological races, VArg1 and VArg2, have been described in Argentina. These races, together with Northern Hemisphere race-1 (NA -1), lack an effector (Ave1) that defines molecular races in tomato and lettuce pathosystems.

The aim of this work is to molecular characterize VArg1, VArg2 and NA -1 isolates using genotyping by sequencing technique ddRAD-seq. For this purpose, in silico digestion simulations of the fungal reference genome were performed and the restriction enzymes MboI and PstI were selected. After sequencing, between 200 - 250 thousand Illumina technology reads were obtained per sample and mapped against the reference genome using the BWA-MEM algorithm. The genome data of sunflower *V dahliae* isolates 85S (France), Vd39 (Germany), S011 and S023 (China) were included in the analysis. Variant calling were determined using the programs Mummer4 and FreeBayes against the reference genome (GCA_000952015.1). The SNP from the seven isolates analyzed were integrated into an 1194 SNP matrix of 126 international isolates for phylogeny analysis using boostraped UPGMA method with the R package poppr.

Results:

In this work, we identified ~ 6000 SNP loci for VArg1 and VArg2, of which more than 100 are unique to each isolate. The analysis of the 1194 SNP matrix allowed us to reproduce the phylogeny previously proposed by several authors and to place the local isolates in the *Verticillium sp* phylogenetic tree. VArg1 and VArg2 were grouped with isolate 85S as a subgroup of the II -1 subclade, while isolate NA -1 was placed together with Vd39 and S011 as a subgroup of clade I. Nucleotide sequence analysis revealed that isolates VArg1 and VArg2 share with 85S a genomic region of approximately 10Kb associated with specific pathogenicity against sunflower, which is not present in NA -1.

Conclusions:

The results obtained strengthen the characterization of *V. dahliae* races affecting sunflower and favor the development of molecular diagnostic methods, epidemiological and surveillance studies, as well as the development of future sunflower breeding strategies for disease control.

Poster Session: http://bit.ly/cab2c-2021-posters
Poster Aguilera et al: https://youtu.be/aS-MqdORk6Q

Unsupervised detection of biologically sound clusters in single-cell expression data

M. Luz Vercesi[1], Ariel Berardino[1], Tomás Vega-Waichman[1], Damiana Giacomini[2], Natalí B. Rasetto[2], Paola Arlotta[3], Alejandro Schinder[2], Ariel Chernomoretz[1,4]

[1]Laboratory of Integrative Systems Biology, Leloir Institute-CONICET, Buenos Aires, [2]Laboratory of Neuronal Plasticity, Leloir Institute-CONICET, Buenos Aires, [3]Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, MA, [4]Phys. Department, School of Sciences, University of Buenos Aires

Abstract

Single-cell RNAseq assays provide a glimpse into cellular transcriptional landscapes. Structures recognized in low-dimensional manifolds probed by this technology allow to explore cell variability, identify new cell-types and uncover developmental pathways at molecular levels. However, detecting patterns at the right scale to unveil relevant biology is not an easy task. Usually, manual curation is needed in order to identify the *right resolution* at which cells should be grouped together. In this preliminary study, we propose an unsupervised method that capitalizes on biological information to uncover biologically meaningful cell clusters. Using our methodology in a public and annotated dataset (Hochgerner2018*) we show that the proposed unsupervised pipeline produces robust results and can recapitulate the annotated cell type provided by the authors of the original publication.

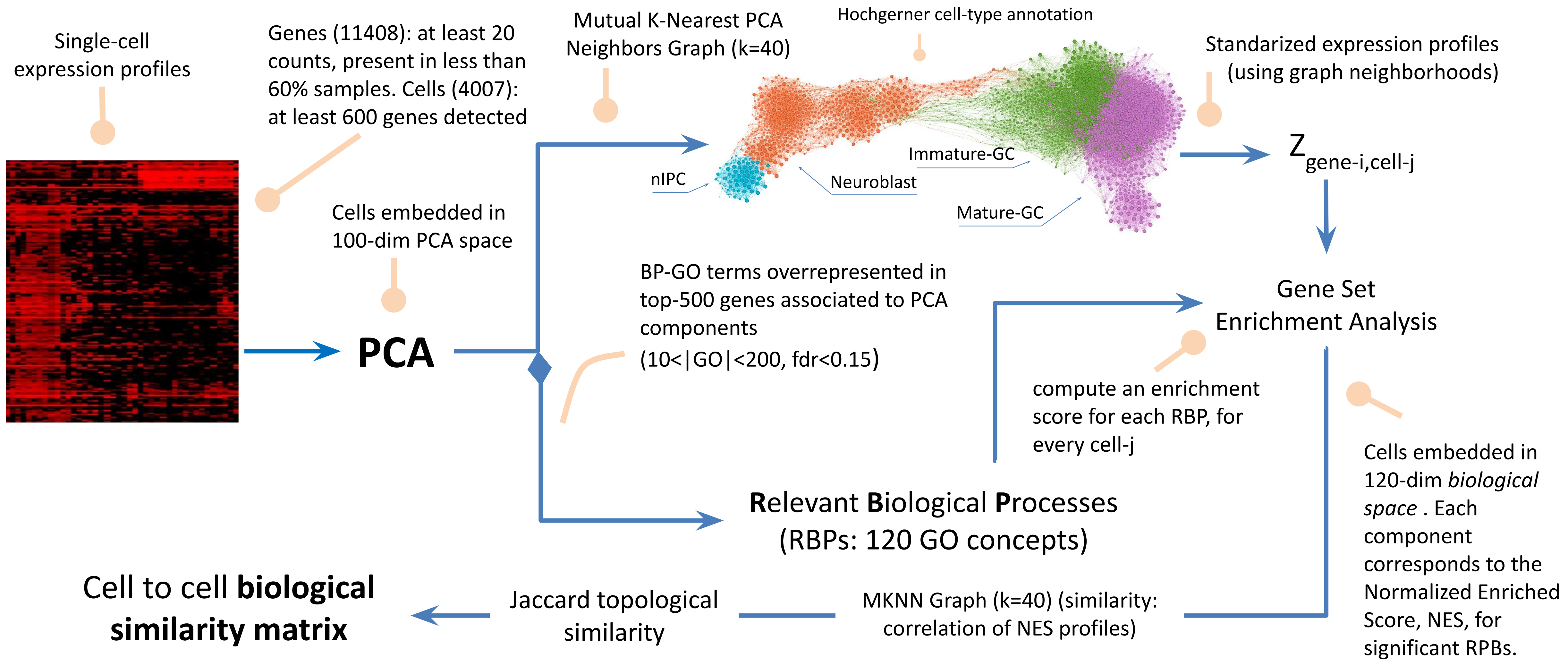
*https://doi.org/10.1038/s41593-017-0056-2

Poster Session: http://bit.ly/cab2c-2021-posters   Poster Vercesi et al: https://youtu.be/JHEa7F8TkPY

The pipeline

- We considered a subset of cells [nIPC (88), Neuroblast (874), immature-GC(1333) and mature-GC(1712)] from the dataset generated at Linnarsson's Lab to study neurogenesis in the dentate gyrus of the hippocampus (Hochgerner2018)
- We built a graph to approximate a low-dimensional manifold to the data and considered three different community detection algorithms to identify cluster structures uncovered by topological similarity patterns.
- We computed a **cell-to-cell biological similarity** matrix and used this information to merge originally detected clusters in a biologically inspired way.
- We evaluated the goodness of the procedure quantifying the mean information carried by the marker-sets of the partition clusters.

Cell biological similarity



Single-cell expression profiles

Genes (11408): at least 20 counts, present in less than 60% samples. Cells (4007): at least 600 genes detected

Cells embedded in 100-dim PCA space

PCA

Mutual K-Nearest PCA Neighbors Graph (k=40)

Hochgerner cell-type annotation

Standarized expression profiles (using graph neighborhoods)

$Z_{\text{gene-i,cell-j}}$

BP-GO terms overrepresented in top-500 genes associated to PCA components (10<|GO|<200, fdr<0.15)

Gene Set Enrichment Analysis

compute an enrichment score for each RBP, for every cell-j

Relevant **B**iological **P**rocesses (RBPs: 120 GO concepts)

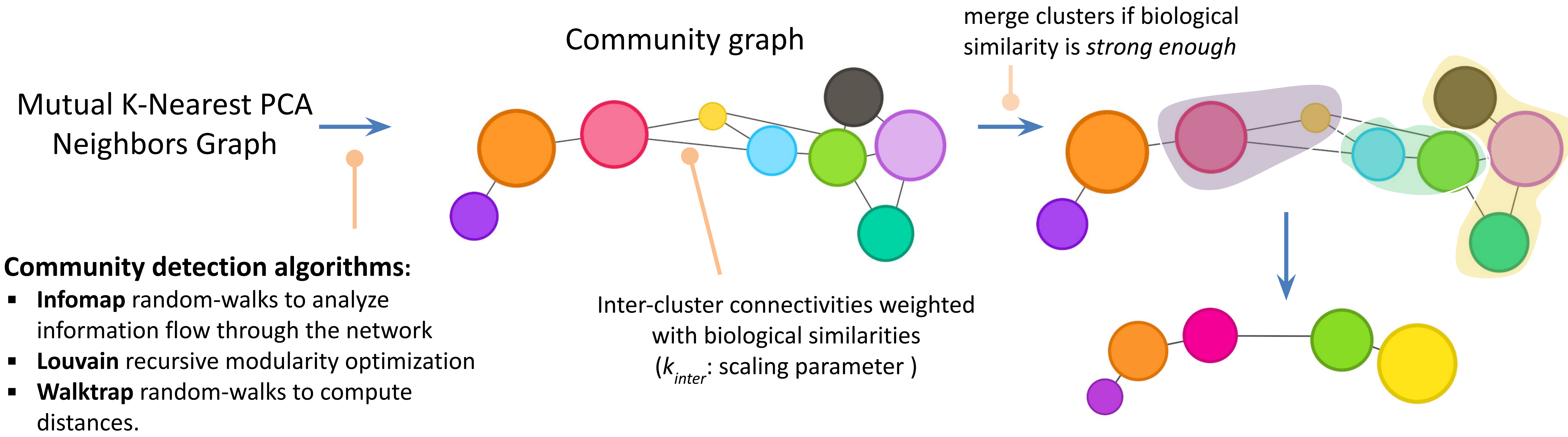
Cells embedded in 120-dim *biological space* . Each component corresponds to the Normalized Enriched Score, NES, for significant RPBs.

MKNN Graph (k=40) (similarity: correlation of NES profiles)

Jaccard topological similarity

Cell to cell **biological similarity matrix**

Biologically inspired cluster merging



Mutual K-Nearest PCA Neighbors Graph

Community graph

Inter-cluster connectivities weighted with biological similarities (k_{inter}: scaling parameter)

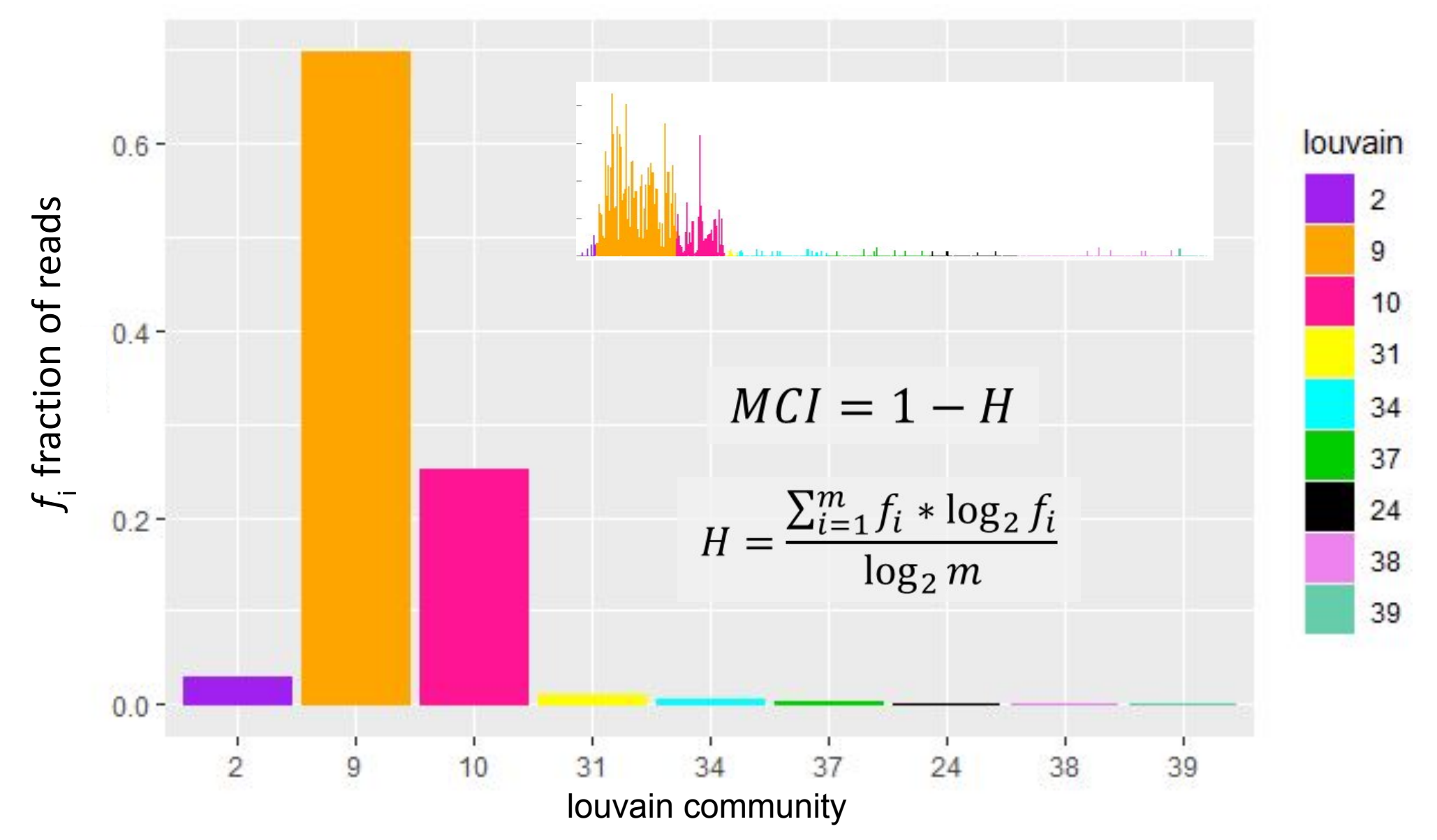
merge clusters if biological similarity is *strong enough*

Community detection algorithms:

- **Infomap** random-walks to analyze information flow through the network
- **Louvain** recursive modularity optimization
- **Walktrap** random-walks to compute distances.

Marker information content

Calb2 - distribution of reads among clusters



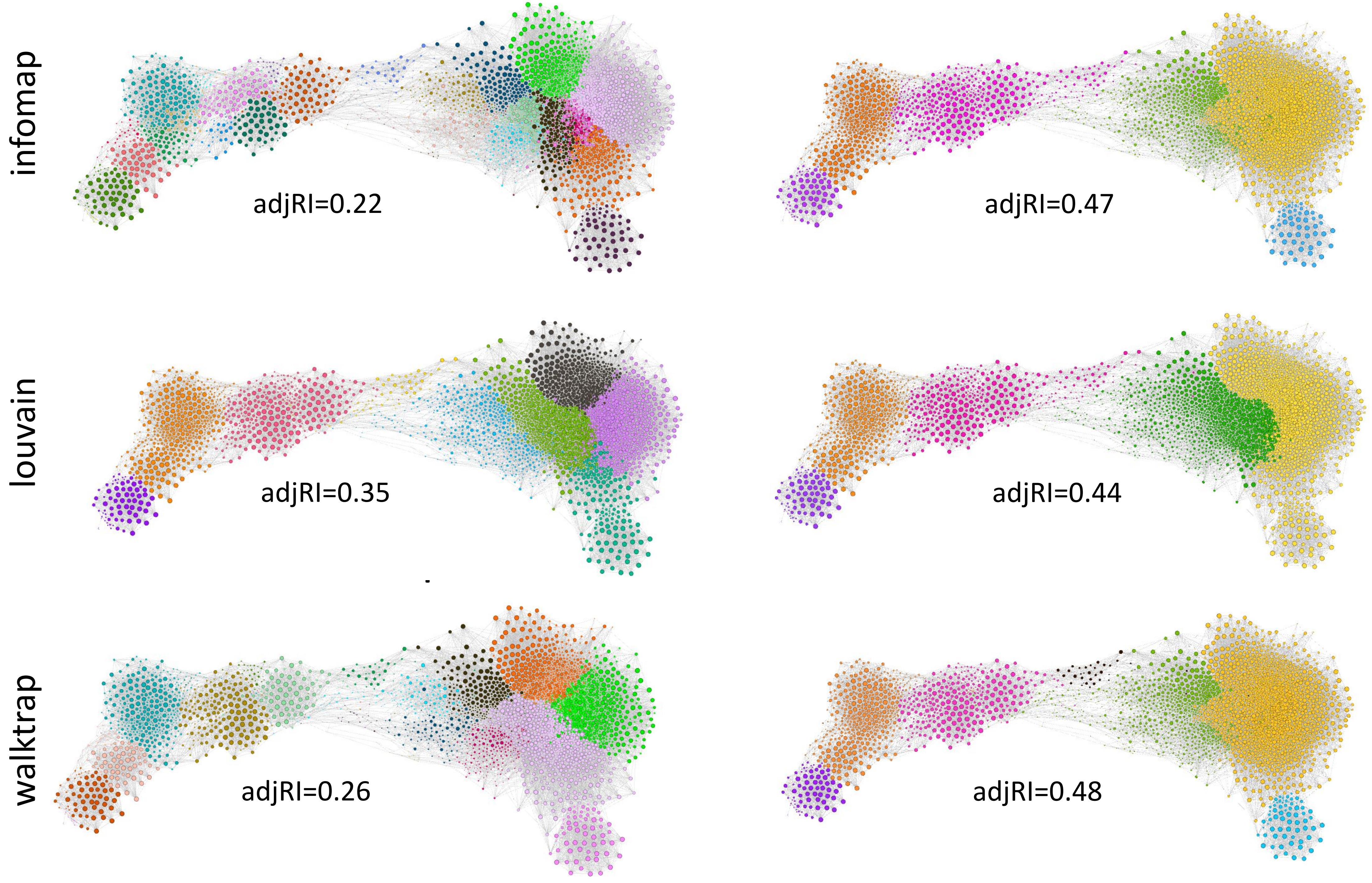
$$MCI = 1 - H$$

$$H = \frac{\sum_{i=1}^{m} f_i * \log_2 f_i}{\log_2 m}$$

The marker carried information measure (MCI) quantifies the inhomogeneity of the marker's read distribution among clusters. A perfect marker minimizes H, the normalized entropy, and get the largest MCI value.

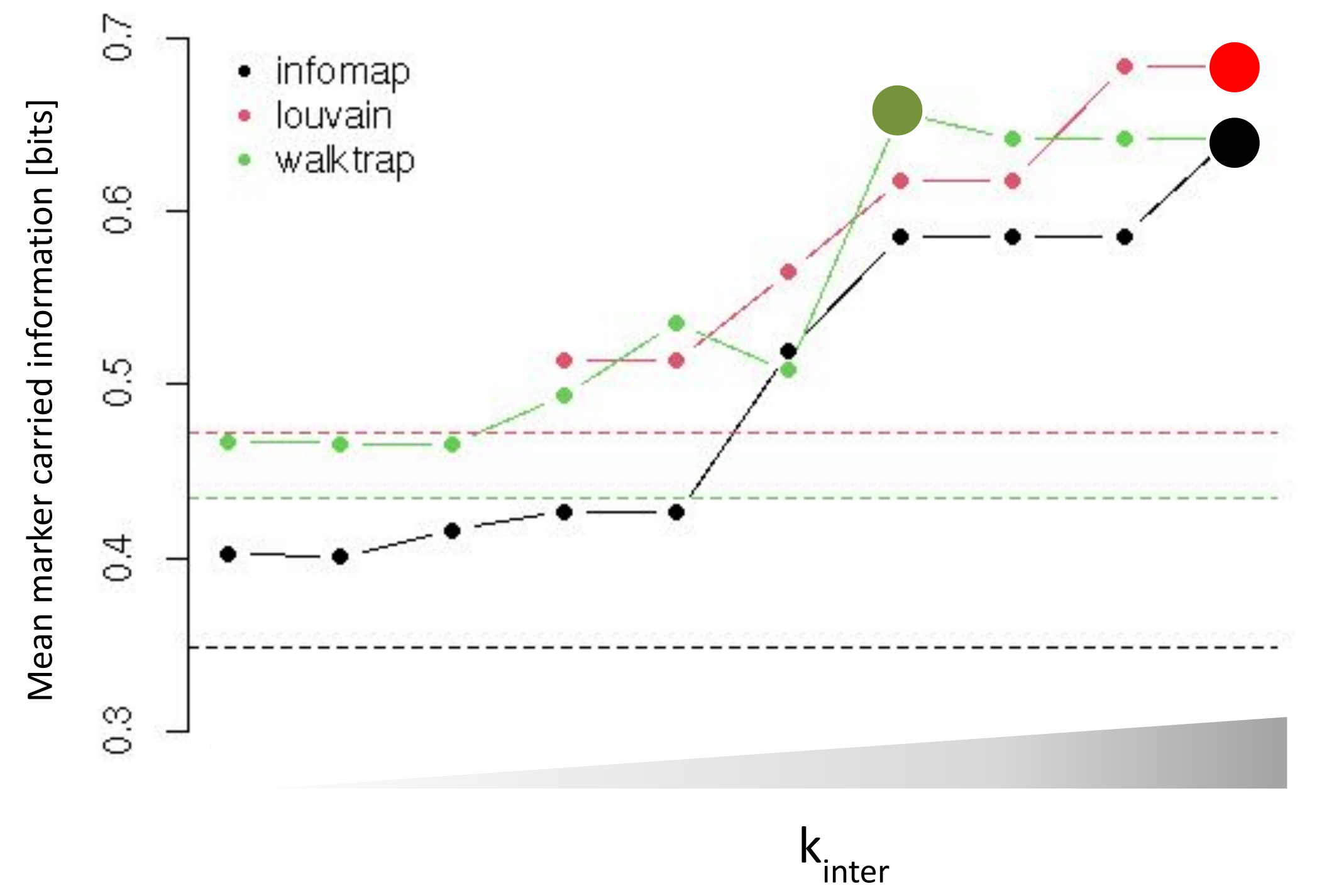
Results

Initial partition   Enhanced clustering



Original and enhanced partitions are displayed in the first and second columns respectively. The adjusted Rand Index wrt the cell-type classification is reported in each case (the larger the value, the better the agreement between partitions) . A noticeable rise of adjRI can be appreciated for the enhanced clusterings (note that, due to inconsistencies in the annotation metadata for cell-types Neuroblasts 1,and Neuroblast 2, we considered a single annotated category: Neuroblasts).

Partition quality assessment



The mean information associated to marker genes is reported for partitions obtained integrating biological knowledge to infomap (black), louvain (red) and walktrap (green) clusters, as a function of the scaling parameter k_{inter}. Large symbols signal the best integrated partition. Horizontal lines correspond to the mean information associated to the original partitions (prior cluster merging).

Conclusions

- We developed a novel unsupervised methodology to identify clusters in scRNA seq datasets. The use of biological information in the coalescence process of seed clusters favors the identification of biologically coherent cell communities.
- Our procedure is robust against the choice of initial clusterings, as long as they provide partitions with comparable resolution. Noticeably, enhanced cell clusters obtained considering different initial partitions can be combined to get more robust results.
- Our approach also identifies putative markers for each detected cell-cluster.

This work was supported by a grant from the National Institute Of Neurological Disorders And Stroke (NINDS) and Fogarty International Center (FIC) (R01NS103758) to P.A. and A.F.S and PICT20183713 grant from ANPyCT to A.C

Reverse vaccinology approach to identify antigens to use in a vaccine against Av. paragallinarum

M. Esperanza Felici [1], Yosef D. Huberman [2], Belkys A. Maletto [1], Rodrigo Quiroga [3].

1. Dpto. de Bioquímica Clínica. Centro de Investigación en Bioquímica Clínica e Inmunología (CIBICI -CONICET). Facultad de Ciencias Químicas. Universidad Nacional de Córdoba. Córdoba, Argentina.
2. Instituto Nacional de Tecnología Agropecuaria (INTA), Estación Experimental Agropecuaria Balcarce. Balcarce, Argentina.
3. Dpto. de Química Teórica y Computacional. Instituto de Investigaciones en Físico-Química de Córdoba (INFIQC-CONICET). Facultad de Ciencias Químicas. Universidad Nacional de Córdoba. Córdoba, Argentina.

Background:

Avibacterium paragallinarum is the causative agent of infectious coryza, an acute disease that affects the upper respiratory system of chickens (*Gallus gallus*). According to Page (1962), this Gram negative bacteria may be classified into three serogroups: A, B and C. This organism is widely distributed in poultry production systems all over the world, causing significant economic losses due to diminished growth performance and egg production. Despite vaccination being the main form of prevention, currently available commercial vaccines, based on inactivated international reference strains, show only partial protection against local strains. In hopes of developing a new vaccine that overcomes this difficulty, antigens with broad protection potential were identified *in silico* from the three Page serogroups of *Av. paragallinarum*.

Results:

To identify antigens with vaccine candidate potential, a reverse vaccinology strategy was followed. Briefly, comparative and subtractive genomics were used in conjunction with sequence analysis to perform a sequential discard of proteins. The first elimination criteria aimed to remove proteins homologous to those of the host to avoid autoimmune reactions, following selection of the essential outer membrane proteins. Afterwards, proteins were classified according to their antigenicity. Furthermore, poorly conserved proteins among the different strains were removed from the dataset. Finally, proteins were characterized regarding their physicochemical properties, secondary structure and signal peptide presence. As a result, proteins deemed the most conserved, essential, antigenic, accessible and non-host homologous, as well as most likely to be easily expressed heterologously, were selected. Among the predicted antigens, type IV pilus biogenesis/stability protein PilW, peptidoglycan-associated lipoprotein Pal, outer membrane protein assembly factor BamE, OmpH family outer membrane protein and rod shape-determining protein MreC were identified as potential candidates qualifying all the set criteria.

Conclusions:

Although only a few strains were analyzed, and *in vitro* and *in vivo* testings are still needed to validate the protective effect of the newly identified antigens, this study provides a basis for the development of a novel subunit vaccine against *Av. paragallinarum*.

Poster Session: http://bit.ly/cab2c-2021-posters
Poster Felici et al: https://youtu.be/z3nd22ZV3bw

Plasmid prediction in *Micrococcus* bacterial strains

Saracho, Hayde; Padilla Franzotti, Carla Luciana; Kurth, Daniel

Planta Piloto de Procesos Industriales Microbiológicos (PROIMI), CCT-CONICET, San Miguel de Tucumán, Argentina. Email: dkurth@conicet.gov.ar

Background: Plasmids are circular or linear extrachromosomal DNA molecules that replicate autonomously and occasionally provide their guests with bacterial extra genetic material important for their survival and adaptation. The sequencing of bacterial genomes has generated a vast wealth of data that can be processed by different computational tools to identify plasmid sequences. This would allow expanding the knowledge about plasmids and their diversity in barely studied bacterial genera such as *Micrococcus*. These are environmental bacteria, and the most known species *M. luteus*, is sometimes associated with skin and opportunistic infections. Other species show potential for biotechnological applications, as they are able to produce antibiotics, biofuels, enzymes and could be applied as biofertilizers or in bioremediation processes.

Results: Draft genomes were obtained from sequencing reads of 20 strains of *Micrococcus.* The combination of different methods on these genomes allowed us to detect the presence of sequences associated with plasmids in 17 of the selected strains. In these sequences, genes directly associated with plasmid functions (replication and segregation) were detected, as well as accessory genes related to resistance to compounds toxins, oxidative stress, and antibiotics.

In order to test the novelty of these predictions, a bipartite bacterial network was constructed with the plasmid predictions and known actinobacterial plasmids. These networks include two types of nodes: “genomic” nodes representing each plasmid or genetic unit, and “protein” nodes representing clusters of protein sequences encoded by the different plasmids. Our network included 833 actinobacterial plasmids, 17 predictions, and 112878 proteins. The network had poor connectivity, with most of the nodes consisting of single elements related to isolated plasmids. From 60615 nodes, 25659 were hypothetical proteins and 41497 included only one protein sequence. From the non-hypothetical proteins, 2138 were annotated as transposases, an abundant element in plasmids, and they formed the largest clusters. This suggests that most actinobacterial plasmids are “unique” and highlights the lack of knowledge on the biology and roles of these mobile genetic elements in Actinobacteria. From a total of 1386 proteins encoded in the plasmid predictions, 915 clusters formed, and 505 of them were exclusively associated with predictions. From these, only 100 were assigned a functional category in the KEGG database, 51 of them encoding proteins associated with genetic information processing and the rest including proteins associated with aminoacids, lipids, energy, and other metabolisms. All categories were already present in the full actinobacterial dataset. Still, this represents a significant addition to the *Micrococcus* plasmid sequences pool.

Conclusions:

Plasmid prediction methods applied to public databases could significantly enrich the known plasmid diversity. Our network analysis allowed to identify the novelty of our predictions in the context of the actinobacterial plasmids. The abundance of hypothetical proteins in the dataset highlights the limited knowledge on plasmid biology, particularly in Actinobacteria.

Poster Session: http://bit.ly/cab2c-2021-posters
Poster Hayde et al: https://youtu.be/4QtJAOWIPfc

IDENTIFICATION OF CALMODULIN-BINDING TRANSCRIPTION ACTIVATOR (CAMTA) GENES IN A NEW VERSION OF THE STRAWBERRY *(Fragaria* x *ananassa*) GENOME

Claudia Rivera-Mora[1], Paz E. Zúñiga[1], Karla Jara-Cornejo[1], Carlos R. Figueroa[1], Lida Fuentes-Viveros[2]

[1] Universidad de Talca, Instituto de Ciencias Biológicas, Talca, Chile

[2]Centro Regional de Estudios en Alimentos Saludables (CREAS), CONICYT-Regional GORE Valparaíso Proyecto R17A10001, Valparaíso, Chile

claudiamarisolrivera@gmail.com

Background:
The commercial strawberry (*Fragaria* x *ananassa* Duch.) is an octoploid hybrid from the spontaneous cross between *Fragaria chiloensis* and *Fragaria virginiana*. The Calmodulin-binding Transcription Activator (CAMTA) is a conserved family of transcription factors found in multicellular eukaryotic organisms. Their transcriptional activity is regulated by calmodulin, the most studied sensor protein and regulator of calcium signaling. In plants, CAMTAs have been identified in different species, and it has been described that they participate in processes related to growth, development and stress tolerance. However, the role of CAMTAs during fruit ripening is unknown. Studies using previous versions of the strawberry genome and peptide database have identified four CAMTA genes. Their expression analysis in vegetative tissues and different fruit stages provide the first information on CAMTAs in strawberry. The publication of a new version of the genome of this species and transcriptomic data in achene and fruit receptacle of different developmental stages offers the possibility of obtaining further information about these genes in strawberry. The main objective of this study is to identify and characterize the CAMTA transcription factor family in the most recent version of the strawberry genome and determine their expression in different tissues by analyzing published transcriptomic data.

Results:
In this study, 13 *FaCAMT*As genes were identified in the *Fragaria* x *ananassa* 'Camarosa' Genome v1.0.a2 and their gene structure determined. Analysis of the deduced protein sequences indicates that they possess the characteristic domains and motifs of CAMTA proteins (CG-1, ankyrin repeats, IPT/TIG and IQ domain) and predictions of subcellular location indicate that they are found in the nucleus. Furthermore, three-dimensional models of CAMTAs proteins show that they have a large number of alpha helices linked by loops and turns. On the other hand, the analysis of achene and receptacle transcriptomes (fruit tissues), leaves and roots, indicate different values of Fragments Per Kilobase of transcript per Million (FPKM) of CAMTAs genes during strawberry fruit ripening.

Conclusions:
This study provides new information that could be useful in future research on the role of CAMTAs during strawberry fruit ripening.

Submission track: FONDECYT/Regular 1201662; ANID BECAS/DOCTORADO NACIONAL 21190862; CONICYT-Regional GORE Valparaíso Project R17A10001 Project; FONDECYT/Regular 1210941

Oral Presentation

Novel structure-based method for linear motif prediction

Gábor Erdős, Zsuzsanna Dosztányi

MTA-ELTE Momentum Bioinformatics Research Group, Hungary

Background

Short Linear Motifs (SLiMs) are compact functional modules that are recognized by specific globular domains. Such interactions play important roles in many biological systems; they regulate the formation of transient protein complexes, direct subcellular localization, and determine the fate of proteins. Although general structure based methods exist for partner prediction, they do not perform well for this specific linear motif mediated systems. Early results from the groundbreaking AlphaFold shows that while in some cases it is able to recognise some linear motif mediated interactions, it is not suitable for partner prediction. Sequence based methods also suffer from various problems, as they cannot capture the global attributes of the binding.

Results

In the presented work we designed a method which is able to identify new binding partners based on the structure and binding mechanism of the complex. Our approach utilizes a unique force-field for each system based on statistical potentials, that maximizes the probability of the given structure compared to a set of random ones.

Conclusion

We successfully applied this method on different systems during testing which gives rise to the possibility to explore the important aspects of molecular recognition and can be applied to systems with known structures where the motif has not yet been discovered.

Oral Presentation

Study of all possible missense variants in the NKX2-5 homeodomain

Jorge Emilio Kolomenski[1], Marisol Delea[2], Leandro Simonetti[3], Liliana Dain[1,2], Alejandro Daniel Nadra[1]

[1] Departamento de Fisiología, Biología Molecular y Celular, Facultad de Ciencias Exactas y Naturales, Instituto de Biociencias, Biotecnología y Biomedicina, Universidad de Buenos Aires, Buenos Aires, Argentina.

[2]Centro Nacional de Genética Médica, ANLIS, Buenos Aires, Argentina.

[3] Department of Chemistry - Biomedical Centre, Uppsala University, Uppsala, Sweden.

Background:

NKX2-5 is a gene coding for a homeobox protein that plays a key role in the formation of the early heart and its function in the adult body and some of its genetic variants (GV) were found to be related with congenital heart disease.

Recently, we compiled, curated and structured GV data for the *NKX2-5* gene from public databases and the scientific literature, obtaining a comprehensive database of GVs. In this study we compiled 143 GVs classified as pathogenic but, for most of them, the molecular mechanisms for their effect is not studied in detail.

In order to gain insight into the molecular biological implications of the GVs, we attempted to predict their effect on NKX2-5 functionality by making a classification of all possible missense GVs in the structured region of NKX2-5 with the help of bioinformatic analysis.

Results:

We developed a BioPython script to generate all the possible missense GVs in the 58 residues of the NKX2-5 homeodomain, which were then classified according to their function and *in silico* studies of stability, change on protein-DNA interaction energy and evolutive conservation.

We proposed a hierarchy of possible effects for the 337 GVs generated, 36 of which were previously classified as pathogenic in our database. With this analysis we found that 80 GVs probably had an effect on protein functionality and predicted a possible effect on functionality for 135 more. In particular, it provided a possible effect on functionality for 29 of the 36 variants classified as pathogenic in our previous study.

Conclusions:

This analysis allows for a guided study of the potential effect of GVs in the structured region of NKX2-5. It helps to understand GVs classified as pathogenic and provides a potential explanation of new ones. We are confident that the pipeline used in this study may help catch up with the constant growth in the number of variants associated with specific phenotypes by providing an initial assessment for unknown variants in other genes.

Poster Session: http://bit.ly/cab2c-2021-posters
Poster Kolomenski et al: https://youtu.be/o9q6OV-o1Yw

“TMBpred”: Versatile server and database as medical oriented tool to estimate immune response in 14 cancer types

Agustina Sofia Torres, Elizabeth Martinez Perez and Cristina Marino Buslje
Fundación Instituto Leloir
agustinasofia.torres@gmail.com, emartinezperez1990@gmail.com, cmb@leloir.org.ar

Background
Cancer is one of the main death causes worldwide cutting across gender and age boundaries. Immune checkpoint blockade (ICB) therapy is becoming a standard-of-care in many malignancies, but only 15–35% of cases derive clinical benefit. Therefore, using biomarkers such as tumor mutation burden (TMB) has become of utmost need to identify patients more likely to respond to ICB.
TMB used to be ascertained by using whole-exome sequencing but it is costly and not feasible in routine clinical settings. Most of the sequencing panels currently used to estimate TMB are inadequately designed and their predictions are inaccurate. Martinez-Perez et. al. (2021) propose cancer-specific panels for 14 malignancies which offer reliable estimates of TMBs and with a reduced number of megabases to be sequenced.
We present a web server and database to offer a precise prediction of TMB in 14 cancer types by using the models published and thus help in the decision of the therapy to be prescribed.

Results
The website allows health professionals to 1) Query the database for the genes to conform the best panel to predict TMB by cancer type. The server will offer 3 strategies differing in the selection of genes to make the models and the predictive performance compared with experimental TMBs, in such a way to make a trade-off between bases to be sequenced (economical cost) and precision. 2) Previous complete or partial information about the genomic alterations, can be input to the server to obtain a TMB prediction.The format of genomic alteration should be supplied in a file with the mutations by gene or the genomic coordinates of each mutation or alternatively a bam file.3) Programmatically access for more intensive and appropriate use by a bioinformatician.

Conclusions
We believe that an easy-to-use server will work as a health service tool colaborating with the discrimination between patients responding and not-responding to ICB therapy, and extending the scope from a published model to the medical practice.

Poster Session: http://bit.ly/cab2c-2021-posters

Development of a web-based platform for taxonomic classification of genomospecies belonging to the *B. cereus* group using machine learning.

Petitti, T.*[1,2], Torres Manno, M.A.[1,2], Cabrera, L.[1,4], Daurelio, L.D.[3], Espariz, M.[1,2]

[1]IPROBYQ-CONICET, Rosario, Argentina; [2]FCByF-UNR, Rosario, Argentina;[3] LIFiBVe, ICiAgro Litoral, UNL, CONICET,FCA, Santa Fe, Argentina; [4] Instituto Politécnico Superior "General San Martín", UNR, Rosario, Argentina .*petittitomas@gmail.com

Background:

The *Bacillus cereus* group is usually categorized into three clades, Clade 1 has pathogenic strains as *Bacillus anthracis*, Clade 2 is composed of *Bacillus cereus sensu stricto*, and *Bacillus thuringiensis*, the former is associated with food poisoning while the latter is used for agronomic purposes for pest control. Clade 3 is the most phylogenetically diverse clade; the strains that compound it have been isolated from very diverse sources. Classification between species within the *B. cereus group* has proven to be very challenging, having reported multiple cases of incorrect classifications or incoherences between taxonomic classification and genomic or phenotypic characteristics. Nevertheless, the correct assignment is of great importance because these assignments are used to predict the performance and safety of bacteria, thus affecting their use for industrial or agronomic purposes.

Results:

In this work, we use the Machine Learning algorithm, Random Forest, to generate classifiers based on gene markers reported for this group. First, 2460 sequences belonging to the Bacillus cereus group were downloaded from GenBank. Of which 2117 were already classified by us, while the remaining 343 were recently uploaded to the database. They were filtered by quality, using parameters of N50, genome size, and the total number of contigs. In this way, 2191 sequences were obtained and validated or reassigned to species using Average Nucleotide Identity (ANI) and multi-locus sequence analysis (MLSA). This resulted in a reassignment of 47.13% of the recently uploaded sequences to the databases. In addition, 5 strains were classified as new genomospecies, named genomospecies 38, 39, 40, 41, and 42.

In order to generate Random Forest-based classifiers, the sequences of 22 gene markers from each of the strains were divided into a training group and a testing group. Of the 2191 sequences, 63 were not included in this analysis because they lacked some gene markers, suggesting that they were incomplete. From the training group, classifier models were generated; their accuracy was evaluated by cross-validation. Thus, it was observed that the classifiers generated to assign the genomospecies of sequences belonging to any of the 3 clades, had an accuracy higher than 98%, being those based on *gyrB*, *pyc* or *lon*, the ones with the highest accuracy. Then, the testing group was used to observe the error of the classifiers, presenting an error of less than 1% for Clades 1 and 2 and less than 4% for Clade 3. Finally, a web-based platform was built to make use of the scripts in a simpler and more user-friendly way.

Conclusions:

The results show that the classifiers generated allow classifying with high accuracy, and the error obtained by using the evaluation group indicates that the models are not over-fitted. The realization of a web platform will make it easier for non-experts in the field to make use of the scripts, reaching a larger number of people. Thus, these classifiers will allow performing mass assignments in metagenomic analysis as well as assignments of new *B. cereus* isolates in a fast and accurate way.

Poster Session: http://bit.ly/cab2c-2021-posters
Poster Petitti et al: https://youtu.be/a7kzvYAHNps

#GTÑ: Introducing the Galaxy Training Ñetwork - Collaboration towards bioinformatics resources in Spanish

Alejandra Escobar[1], Maria Trinidad Bernardi[2], Patricia Carvajal-López[1] and Wendi Bacon[3] [1]EMBL-EBI
[2]IQUIBICEN CONICET
[3]The Open University, United Kingdom
ales@ebi.ac.uk

Background

Short courses for up-skilling wet-lab scientists in bioinformatics are key, both for today's and tomorrow's scientists. Bioinformatics infrastructure is expensive, which can be solved through the use of free compute infrastructure offered by the Galaxy Project. Still, a notoriously steep learning curve often prevents scientists from pursuing bioinformatics. Materials provided only in English exacerbate this problem. However, most scientific research is done in English (like this conference!), so is English the preferred medium? If not, how can we automate translation - is GoogleTranslate sufficient?

Questions

Which is best for bioinformatics learning - English, Google-Translated Spanish, or Human-Translated Spanish for native Spanish speakers?

Methods

Native Spanish-speaking bioinformaticians from Mexico, Spain, and Argentina have created translations for one workshop containing:

- three Galaxy tutorials
- two lecture slide decks
- three walkthrough videos worth of subtitles, to be dubbed

We will run this free virtual workshop for native Spanish speakers, randomly streaming across the three language versions of materials, and assessing learning experience and perception through surveys.

Results

We identified multiple challenges in translation.
Low availability of peer-reviewed materials or consensus in translated bioinformatics Spanish terminology
Significant time required for maintaining and building translations
Spanglish Markdown environment requires semi-translation - i.e. tools, parameters, and background code must remain in English.
Professional translators are unaccustomed to markdown format.

Future

We will assess the trainer perspective, analysing the language effect on usefulness for local, native Spanish-speaking trainers in undergraduate environments. We'll also build a pathway for scientists to create material in Spanish, necessitating a Spanish peer review task force. We will undertake a larger project on the art of bioinformatics semi-automated translation to roll out translation across the Galaxy Training Network. Translators, tutorial guinea pigs, workshop participants and social media sharers are needed! Help! [@canicamxli #GTÑ]

Poster Session: http://bit.ly/cab2c-2021-posters
Poster Escobar et al: https://youtu.be/DHo11hbTRGo

Uncovering wheat growth promotion traits by inspecting PGPB genomes

Torres Manno MA[1], Gizzi FO[2], Blancato VS[2], Daurelio LD[3], Espariz M[1,4]

[1]FCByF-UNR;[2]IBR-CONICET-UNR; [3]LIFiBVe, ICiAgro Litoral, UNL; [4]IPROBYQ-CONICET-UNR

Background:

Wheat is one of the principal cereals of Argentine agriculture and its cultivation is considered strategic in rotations due to its contribution to the sustainability of the soils. Multiple factors can affect the wheat production generating considerable losses in its yield. Plant growth promoting bacteria (PGPB) can colonize the rhizospheres of plants, and act as biofertilizers and antagonists of pathogens (biopesticides). Due to this, they emerged as a technological alternative for a sustainable agricultural exploitation, as a replacement for agrochemicals. Many of these microorganisms belong to the genus *Bacillus* and proliferate in soils exploited agriculturally and its mechanism are yet to be uncover.

Results:

In this work we characterize six wheat associated strains presenting PGPB and biocontrol properties belonging to *Bacillus velezensis* and *Priestia megaterium* (formerly known as *Bacillus megaterium*). The whole genome sequences were determined using Illumina and PacBio technology. The taxonomy identity was defined comparing available genomes from 478 *B. velezensis* and 113 *P. megaterium* group. The 591 available strains along with the 6 isolated were complete reclassified using Multiple Locus Sequences Analysis (MLSA) and Average Nucleotide Identity (ANI). A comparative genomic analysis was processed in order to identify the plant growth promoting mechanism of these strains. Known secondary metabolite and general PGP pathways were searched first using the GeM-Pro algorithm. This pathway search upon the six strains and available genomes exposes some of the possible mechanisms in growth promoting and biocontrol. Additional potential pathways were searched using the antiSMASH platform resulting in potential new pathways for *P. megaterium* and *B. velezensis* isolated strains. Another comparative genomic analysis with these new pathways was performed with the available genomes with the aim of finding the exclusive genes that correspond with the differential plant growth promoting phenotypes. As result, we found exclusive pathways in the *P. megaterium* strains involving Non-Ribosomal Peptide Synthases (NRPS) and Polyketide Synthase (PKS) that were not detected in the non-redundant nucleotide GenBank database. Secondly, thanks to the PacBio technology, we confirm that these gene clusters are coded in two different plasmids. Furthermore, a rare NRPS pathway were found in the two *B. velezensis* strains that were not common in the analyzed genomes from *B. velezensis* group.

Conclusions:

The results show that there are 33% misclassified strains of *B. velezensis* group when using complete genome tools. This implies the necessity of a framework for true genomoespecies classification. The new pathways found in the isolated strains may suggest that these are recently acquired gene clusters as result of adaptation to the environment.

Poster Session: http://bit.ly/cab2c-2021-posters
Poster Torres Manno et al: https://youtu.be/nTL0cW0fo0g

Usefulness of *in silico* prediction tools to confirm correlation between phenotype and genotype in type 2 VWD

Adriana I Woods PhD[a], Débora M Primrose PhD[b], Juvenal Paiva MS[c], Analia Sánchez-Luceros MD, PhD[ac]

[a]Laboratory of Hemostasis and Thrombosis, IMEX-CONICET-National Academy of Medicine. Buenos Aires City, Argentina

[b]Higher School of Engineering, Informatics and Agri-food Sciences, University of Morón. Buenos Aires, Argentina.

[c]Department of Hemostasis and Thrombosis, Hematological Research Institute, National Academy of Medicine. Buenos Aires City, Argentina

Background:

Type 2A, 2M and 2B variants of von Willebrand disease (VWD) are characterized by qualitative defects, with dominant inheritance. Determining whether variants observed in these patients are related to their clinical and laboratory phenotypes requires additional experimental approaches, which are expensive, laborious and time-consuming.A large variety of in silico prediction tools have been designed to evaluate the prediction of pathogenic potential of genetic variants on the structure and/or function of theresulting protein. However, their performance can vary markedly. Using genotypic variants of known pathogenicity related to type 2A, 2M and 2BVWD, we usedseveral prediction tools and algorithms to assess the percentage of correlation between the phenotype-genotype according to the resulting pathogenicity for each genotypic variant, with the purpose of evaluating their use as an alternative, efficient and reliable tool, mainly for characterizing novel genetic variants. The following in silico tools were used: PolyPhen-2; SIFT; SIFT4G, Revel, Mutation Assessor, MutPred, Panter; MutationTaster; Meta-SNP; CADD; SNP & GO; FATHMM; PhD-SNP; Provean; BDGP; HSF; ASSP, I-Mutant.

Results:

The performance of commonly used in silico methods resulted highly positive.We have found a high percentage of correlation between the different in silico methods and the genotypic variants studied. The prediction scores of type 2A, 2M, and 2B VWD variants showed correlations with the degree of functional defects and the severity of clinical phenotypes.

Conclusion:

Given the high phenotype-genotype correlation achieved, in silico prediction methods might be an excellent tool for supporting the classification of genotypic variants related to VWD, especially those novel variants.

Poster Session: http://bit.ly/cab2c-2021-posters
Poster Woods et al: https://youtu.be/ZiCbzCw1Qfw

Development of sex-specific markers in Piaractus mesopotamicus

Florencia C. Mascali[1,2], Victoria M. Posner[1], Emanuel A. Romero Marano[1], Felipe Del Pazo[1,2],
Miguel Herminda[3], Paulino Martinez[3], Juan A. Rubiolo[2,3] and G. Vanina Villanova[1,2]
[1]LMBA-FCByF-UNR- MPCyTSF - Rosario – Argentina
[2]CONICET CCT-Rosario- Rosario- Argentina
[3]Facultad de Veterinaria - USC - Lugo – España
flomascali@gmail.com

Background

Even though sex is nearly universal to eukaryotic life, sex-determining mechanisms are diverse and can evolve rapidly. In fish, it had been described from environmental to genetic sex determinations systems, with one o more major genes or genomic regions related to sex. The ability to determine sex is one of the first applications in the development of genomic resources. In aquaculture, it is important because economically valuable traits may be related to sex.

Piaractus mesopotamicus, popularly known as Pacú, is a freshwater neotropical native fish. It is of major importance for South American aquaculture. Despite its commercial importance, there is little genomic information about this fish. Production and market conditions indicate that a genetic improvement plan for Pacú could have a positive impact on production. Sex-associated markers are useful in this context for precocious sex identification, especially in pacú that lacks both sex chromosomes and sexual dimorphism.

Results

To identify markers associated with the binary sex trait, a genome-wide association analysis (GWAS) was conducted over 133 individual fishes (67 females and 66 males) of Pacú with both TASSEL and PLINK programs. A male genome previously assembled was used as the reference genome. It was called 21786966 variable sites, of which 12481745 were biallelic and passed the filters respectively. 465 variable sites (SNPs and indels) that passed a Bonferroni correction limit (α=0.001) were preselected as primary candidate markers Those sites were ordered by p-value and manually analyzed looking for those sites that presented a robust heterozygosis difference between sexes, with all alternative homozygous in females, and heterozygous or reference homozygous in males. The top 5 contigs with variable sites were used to design PCR primers for putative sex-specific marker region identification. In addition, reads from 4 females and 4 males were aligned on the 5 contigs. It was observed that 2 of them have regions present only in male fish. Seventeen sets of primers were designed to span across the regions of interest. Amplification of each putative marker was standardized using DNA of one male and one female fish. SNPs were confirmed by sequencing, and three PCR presented differential bands between male and female samples. Finally, three markers were selected and validated in a higher number of samples.

Conclusions

In conclusion, we developed and validated a group of sex-specific primers that allow identifying sex in Pacú.

Poster Session: http://bit.ly/cab2c-2021-posters
Poster Mascali et al: https://youtu.be/Saj6yupLqlM

Exploring the conformational diversity of RNAs with CoDNaS-RNA

Martín González Buitrón[1], Ronaldo Romario Tunque Cahui[2], Emilio García Ríos[2], Layla Hirsh[2], Gustavo Parisi[1], María Silvina Fornasari[1], Nicolás Palopoli[1]

[1] Departamento de Ciencia y Tecnología, Universidad Nacional de Quilmes - CONICET, Buenos Aires, Argentina

[2] Departamento de Ingeniería, Pontificia Universidad Católica del Perú, Lima, Perú

Background:

The structural dynamics of biomolecules are essential to hold their functions. In particular, the conformational changes in native ensembles of RNAs provide a good description of their native states[1]. A rugged energy landscape model is associated with the nature of their conformational changes relevant for function[2] but is also triggered by external factors such as temperature or pH changes and their interactions with other molecules[3].

We have developed CoDNaS-RNA, a database of conformational diversity in RNAs that helps researchers gain insights of the relationship between RNA dynamics and function in known and novel examples of well-established biological relevance.

Results:

Each entry in CoDNaS-RNA compiles a cluster of known structures of RNAs with the same sequence. Pairwise structural comparisons allow an extended description of conformational diversity, including the visual and quantitative exploration of structural variability and the recent addition of global and cluster-based heatmaps representing the hierarchical clustering of conformers. Additional annotations about structural features, molecular interactions and biological function are provided. These features showcase possible associations between the diversity in the native ensemble and physicochemical, biological or functional modulators, such as pH, temperature and ligand presence, with 66% of CoDNaS-RNA clusters having an RNA-protein interaction.

CoDNaS-RNA provides extended data on cases of interest such as the HIV trans-activation response (TAR) element, which displays a stem-loop arrangement[4]. Its interaction with the viral protein Tat is unique to HIV-1 and thus an ideal target for drug design. This interaction is associated with a conformational change evidenced by two independently solved TAR conformers. While their secondary structure representations are identical, exploration of their tertiary structures in CoDNaS-RNA shows that TAR undergoes large displacements (RMSD=4.24 Å). This flexibility of TAR could help explain different responses to drugs arising from unique interactions imposed by the alternative conformers.

Conclusions:

CoDNaS-RNA constitutes the first RNA database integrating structural data and annotations to highlight conformational diversity as an essential feature for understanding RNA dynamics and function. The information recovered from the database is useful to develop and explore hypotheses on the dynamic-structure relationship of RNAs. The latest version of CoDNaS-RNA and all associated data are freely available at http://ufq.unq.edu.ar/codnasrna

References:

[1] Ganser et al., Nat Rev Mol Cell Biol, 2019. PMID: 31182864

[2] Simon and Gehrke, Biochim Biophys Acta, 2009. PMID: 19501200

[3] Mustoe et al., Annu Rev Biochem, 2014. PMID: 24606137

[4] Aboul-ela F. et al., Nucleic Acids Res, 1996. PMID: 8918800

Poster Session: http://bit.ly/cab2c-2021-posters
Poster Gonzales Buitron et al: https://youtu.be/s6U9QhvyZac

Bioinformatics approach to reveal the role of alternative spliced genes in potential colorectal cancer subtypes

Marcos Zornn[1*], Eliana Vallmitjana[1*], Gabriela A. Merino[1,2,3].

[1]Facultad de Ingeniería, Universidad Nacional de Entre Ríos (FI-UNER), Oro Verde, Argentina.
[2] Instituto de Investigación y Desarrollo en Bioingeniería y Bioinformática (IBB)-CONICET-UNER, Oro Verde, Argentina.
[3] Instituto de Investigación en Señales, Sistemas e Inteligencia Computacional (sinc(i))-CONICET-UNL, Santa Fe, Argentina
*Both authors contributed equally to the present work

Background:

Colorectal cancer (CRC) is the third most common cancer and the second most deadly cancer worldwide. Therefore, it is important to develop novel strategies for understanding this disease aiming to reduce morbidity and mortality in the future by performing specific therapies. Several community efforts have been carried out to characterize the transcriptomics CRC subtypes. However, although alternative splicing has been reported to be involved in the tumorigenesis and development of CRC, it has not been deeply analyzed jointly to CRC subtypes.

Results:

Here we present a bioinformatics approach focused on the role of alternatively spliced genes in potential CRC subtypes, predicted by unsupervised clustering strategies. Genes and isoforms expression data from next-generation sequencing technologies were obtained from TCGA of the Colon Adenocarcinoma and the Rectum Adenocarcinoma projects (640 samples). Tumoral samples were randomly divided to generate 12 independent subsets, due to the imbalance between tumoral (591) and normal (49) samples. Differential expression analyses for each independent subset were carried out, and results were then compared. A total of 2,686 differentially expressed genes (DEG) and 524 differentially spliced genes (DSG) between tumoral and normal samples were obtained. The k-means algorithm was used for obtaining sample clusters by using DEG and DSG. Then, a consensus was made identifying eight potential CRC subtypes. Representative DSGs characterizing each subtype were found by analyzing splicing patterns that differentiated it from the others and from the normal group. Functional annotations of representative DSGs were used for functional characterization of each potential CRC subtype.

Conclusions:

Unsupervised learning and differential expression analyses can be successfully used to discover and characterize potential CRC subtypes and alternatively spliced genes and set a basis for potential disease treatment strategies.

Poster Session: http://bit.ly/cab2c-2021-posters
Poster Zornn et al: https://youtu.be/Q75gZt9tT5Q

Simulation of the endocytic and secretory pathways using Agent Based Model

Franco Nieto, Ignacio Cebrian and Luis S. Mayorga
IHEM (UNCuyo, CONICET), Facultad de Ciencias Exactas y Naturales, UNCuyo, Mendoza

Background:

Intracellular traffic is a central process in the cellular physiology. Numerous macromolecules must be transported in the endocytic and exocytic pathways for the correct function of a eukaryotic cell. However, the way by which macromolecules are transported between compartments is still a matter of intense debate. Intracellular transport occurs in dynamic organelles that merge, divide, and change position and shape, while altering their composition by complex networks of molecular interactions and chemical reactions. We are interested in cross presentation in dendritic cells. These cells belong to the antigen-presenting cells group, and play a central role in linking innate sensing of pathogens and antigen processing to adaptive immune responses. Antigen cross-presentation consists of several steps. First, the antigenic proteins are internalized by endocytosis. Then, by still not well-characterized mechanism, these molecules are translocated to the cytosol for proteasome degradation. The resulting peptides are transported back to the phagosome and loaded onto Major Histocompatibility Class I complexes (MHC-I). Finally, the MHC-I/peptide complexes migrate to the plasma membrane to trigger a CD8+ T lymphocyte cytotoxic response.

Results:

Our group has developed a simulation of the endocytic pathway (early, late, sorting and recycling endosomes), based on a combination of agent-base modeling and ordinary differential equations. To simulate cross-presentation, we incorporated a secretory pathway, including the ER compartment, ERGIC, three Golgi cisternae and a Trans Golgi Network. The model also considered plasma membrane and cytosolic compartments. This complex endomembrane system was able to reproduce antigen internalization, translocation to the cytosol, processing to peptides, incorporation to membrane-bound organelles and transport to the plasma membrane. The parameters of the simulations were adjusted by experimental results; nevertheless, more empirical data will be required to specify the parameter for each step of cross presentation. All compartments preserved their identity and functionality for more than 300.000 ticks (equivalent to 5 hs. of cellular life). This results show that an endomembrane system able to reproduce complex processes that strongly depends on cargo trafficking can be organized by the interaction of individual agents following simple rules.

Conclusions:

This modeling strategy successfully reproduce transport in the endocytic and secretory pathways and the slow appearance of MHC-I complexes loaded with peptide on the cell surface.
We expect that the active dialogue between simulations and experimental results will foster our understanding of the logic underlying the transport mechanisms that efficiently sort a large number of macromolecules to their final destination inside and outside the cell.

Poster Session: http://bit.ly/cab2c-2021-posters
Poster Nieto et al: https://youtu.be/NH77muHDZgk

Integrating RNA-Seq data to boost peptide-MHC class I binding predictions

Heli Magalí García Álvarez[1] and Morten Nielsen[1,2]

[1] IIBIO, UNSAM-CONICET, Buenos Aires, Argentina, [2] Department of Health Technology, Technical University of Denmark, Lyngby, Denmark

Background:

The presentation of antigens by the class I Major Histocompatibility Complex (MHCI) is a crucial event for adaptive immunity. Peptides bound to MHCI in the cell surface can interact with T-cell receptors (TCR) in CD8+ lymphocytes. Peptide-MHCI:TCR binding is the first signal required to trigger cytotoxic T-cell activation which finally leads to the eradication of virus or parasite-infected cells. Moreover, this already mentioned mechanism also plays a key role in tumor immune surveillance, allowing for the elimination of cancer cells that present non-self or mutated peptides in the MHCI context. The MHCI processing and presentation pathway consists of several steps, many of which can be predicted using computational methods. Nevertheless, the majority of the methods available nowadays do not take into account the relative abundance of proteins in cells, or transcripts that give rise to them, when assessing which peptides can be presented by MHCI.

Objective, methodology and results:

In this work we integrated gene expression data to our state-of-the-art method to predict peptide-MHCI binding: NetMHCpan-4.1. To accomplish this objective we extracted the frequencies of protein coding mRNAs from RNA-Seq experiments performed on different human cell lines and tissue samples, which were in turn assayed by liquid chromatography-mass spectrometry (LC-MS) techniques to determine their MHCI ligands. In this way, the MHCI ligands in our training dataset were mapped onto their source proteins, and subsequently mRNAs, and were given a relative abundance value. The present work shows that including this new feature boosts peptide-MHCI binding predictions.

As regards the methodology, we employed a modified version of the fully-connected neural network NNAlign_MA, upon which NetMHCpan-4.1 is also built. For the training of this neural network we used the data generated by the aforementioned immunopeptidomics experiments (LC-MS assays) with paired or unpaired gene expression levels. In the cases where no RNA-Seq experiments were available for a certain sample, we used a reference RNA-Seq dataset from the Human Protein Atlas (HPA). The pan-specific models were trained on a five-fold cross validation scheme, as well as with independent data (large set of neoepitopes and LC-MS ligands). The most striking results of this study show that, not only in cross-validation but also in the evaluation with independent data, the models trained with this new feature perform significantly better than the models trained without it. Indeed, if the models are trained only using the external HPA data as a proxy for gene expression, we also find that they significantly outperform models trained without gene expression and NetMHCpan-4.1.

Conclusions:

In conclusion, our work highlights the relevance of including the mRNA expression level of a peptide's source protein to improve MHCI ligand and neoepitope predictions. In sum, the experimental evidence and the models derived from this research show that there is a balance between a peptide's abundance and its MHCI binding affinity, which implies that highly abundant peptides that are considered weak binders, or even non-binders, can also potentially bind to MHCI.

Poster Session: http://bit.ly/cab2c-2021-posters
Poster Magalí García Álvarez et al: https://youtu.be/TRkVeZWSO5Q

K-nearest neighbor correction of confounding effects in gene expression measurements

Macarena Alonso[1]; Cristina Marino-Buslje[1]; Saikat Banerjee[2]; Johannes Soeding[3]; Franco L Simonetti[1]

[1]Bioinformatics Unit, Fundación Instituto Leloir, Buenos Aires, Argentina

[2]Department of Statistics, University of Chicago, Chicago, IL, USA

[3]Quantitative and Computational Biolog, Max Planck Institute for Biophysical Chemistry, Goettingen, Germany

Background. Gene expression measurements can be dominated by strong confounding effects such as technical details of RNA recovery, sample conservation, sequencing, environmental and biological factors (patient's age, gender, diseases and many others). Computational methods rely on a good quality control of these confounders to be able to provide reliable results. Typical tools for removing confounding effects from gene expression data involve using a linear regression model to regress out known confounders. Statistical tools such as Principal Components Analysis (PCA) or PEER analysis can infer new covariates that can be later regressed out in the same way as with known confounders.
Our objective is to improve confounder correction in order to obtain better gene expression prediction models. If the systematic noise can be controlled and minimized, this could lead to an upgraded model generalization.
We developed a new method based on a k-nearest neighbour (KNN) approach where we assume that confounding effects dominate the gene expression. If the samples are close to one another in the expression space, we expect them to be close to one another in the confounder space and hence, to be able to correct them.

Methods. We performed gene expression corrections (i) using PCA components from sample genotypes, (ii) calculated inferred PEER covariates and (iii) applied our KNN method with different parameter combinations. Next, we used PrediXcan to build a multiple linear regression model of each gene expression levels using SNPs in a 1Mb window around each gene. The GEUVADIS dataset was used to learn/train a model for each corrected gene expression from sample genotypes using PrediXcan. We tested the models' accuracy by predicting gene expression levels on the GTEx dataset. GEUVADIS and GTEx contain data from genotyped samples and RNA-seq gene expression measurements. Both contain expression levels in lymphoblastoid cell lines (LCL).

Results. The accuracy was calculated using the Pearson correlation coefficient (PCC) between the predicted and measured gene expression. For every confounder correction setting we compared the accuracy in prediction. Our initial results indicate that KNN correction provides comparable results to PEER. The elastic-net regression used by PrediXcan also finds more gene models with explanatory SNPs ($\beta \neq 0$) compared to using PEER alone. An analysis of the KNN correction term indicates that it correlates well with known confounders obtained from phenotype data from GTEx samples.

Conclusions. We introduced a method for correcting gene expression measurements that was validated with a different dataset and compared with other existing methods. We plan on looking for more adequate parameters and which combination of the methods gives a better result.
We hope that our work can contribute to the development of new methods for correcting gene expression.

Oral Presentation

Computational and Structural Biotechnology Journal 19 (2021) 3964–3977

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Insight into membraneless organelles and their associated proteins: Drivers, Clients and Regulators

Fernando Orti [a], Alvaro M. Navarro [a], Andres Rabinovich [a], Shoshana J. Wodak [b,c,1], Cristina Marino-Buslje [a,1,*]

[a] *Bioinformatics Unit, Fundación Instituto Leloir. Avda. Patricias Argentinas 435, Buenos Aires B1405WE, Argentina*
[b] *VIB-VUB Center for Structural Biology, Flemish Institute for Biotechnology, Brussels, Belgium*
[c] *Structural Biology Brussels, Vrije Universiteit Brussel, Brussels, Belgium*

ABSTRACT

In recent years, attention has been devoted to proteins forming immiscible liquid phases within the liquid intracellular medium, commonly referred to as membraneless organelles (MLO). These organelles enable the spatiotemporal associations of cellular components that exchange dynamically with the cellular milieu.

The dysregulation of these liquid–liquid phase separation processes (LLPS) may cause various diseases including neurodegenerative pathologies and cancer, among others.

Until very recently, databases containing information on proteins forming MLOs, as well as tools and resources facilitating their analysis, were missing. This has recently changed with the publication of 4 databases that focus on different types of experiments, sets of proteins, inclusion criteria, and levels of annotation or curation.

In this study we integrate and analyze the information across these databases, complement their records, and produce a consolidated set of proteins that enables the investigation of the LLPS phenomenon. To gain insight into the features that characterize different types of MLOs and the roles of their associated proteins, they were grouped into categories: High Confidence MLO associated (including Drivers and reviewed proteins), Potential Clients and Regulators, according to their annotated functions. We show that none of the databases taken alone covers the data sufficiently to enable meaningful analysis, validating our integration effort as essential for gaining better understanding of phase separation and laying the foundations for the discovery of new proteins potentially involved in this important cellular process.

Lastly, we developed a server, enabling customized selections of different sets of proteins based on MLO location, database, disorder content, among other attributes (https://mlos.leloir.org.ar).

Poster Session: http://bit.ly/cab2c-2021-posters
Poster Marino-Buslje et al: https://youtu.be/sR3haHFBrik

[*] Corresponding author.
E-mail address: cmb@leloir.org.ar (C. Marino-Buslje).
[1] Contributed equally

https://doi.org/10.1016/j.csbj.2021.06.042

How accurate are AlphaFold2 predictions for close homologous proteins?

Cristian Emanuel Guisande Donadio[1,2], Juan Mac Donagh[1], Nicolas Palopoli[1,2], Maria Silvina Fornasari[1,2] and Gustavo Parisi[1,2]

[1] Departamento de Ciencia y Tecnología, Universidad Nacional de Quilmes, CONICET, Bernal, Buenos Aires, Argentina.
[2] Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Argentina

Background:

In the last year the computational tool AlphaFold2, developed by DeepMind, reached an impressive performance in the prediction of protein structures, with an accuracy similar to experimental techniques. AlphaFold2 is based on a novel neural network architecture that attends over evolutionary information, codified in a multiple sequence alignment (MSA), to create a novel representation of the sequence and the position-relative distances for the next step. Besides this outstanding breakthrough of AlphaFold2 in predicting protein 3D models, new questions appeared and still remain unanswered. This work focuses on the AlphaFold capacity to predict 3D models for close homologous proteins. Our hypothesis sustains that the evolutionary information contained in an MSA is not diverse enough to correctly predict close related sequences. Consequently, it is expected that the quality of AlphaFold2 predictions for a pair of evolutionary related sequences would decrease with sequence similarity.

Results:

To address this hypothesis, we sequentially clustered the recently released AlphaFold models for the complete human proteome deposited in the EMBL-EBI database (https://alphafold.ebi.ac.uk/). Clustering was performed with CD-HIT at 60% identity and 60% coverage resulting in 758 clusters of putative paralogs. For each protein we mapped the presence of a corresponding crystallographic structure deposited in PDB. Structural alignments between each model and its corresponding PDB structure (where available) were obtained using TM-Align. The resulting RMSDs were taken as errors in the predictions made. For each cluster of homologous proteins obtained before and for each pair of proteins in each cluster, we estimated the average error and average identity percentage between proteins. The average error between two proteins was normalized by the average structural divergence between AlphaFold models and PDB structures for each pair of proteins. We have found that the error in the estimation of 3D models using AlphaFold2 abruptly increases for close homologous pairs of proteins (above ~90% sequence identity).

Conclusions:

This finding could be explained as a limitation of AlphaFold2 to model too close homologous proteins when the evolutionary information in the input alignments for AlphaFold2 is extremely similar. Our error estimation as a function of protein divergence could be used to confidently predict 3D models of homologous proteins below a given cutoff.

Poster Session: http://bit.ly/cab2c-2021-posters
Poster Guisande Donadio et al: https://youtu.be/s2A_kUa2JnY

Genome-wide analysis of *Trichoderma harzianum* functional evolutionary adaptation

Maria Belén Aguer[1] and Gustavo Parisi[2]
1 Universidad Nacional del Noroeste de la Provincia de Buenos Aires
2 Departamento de Ciencia y Tecnología, Universidad Nacional de Quilmes, CONICET, Bernal, Buenos Aires, Argentina.

Background:

Trichoderma species are among the most widely used microbial biological control agents in agriculture. *Trichoderma* is a filamentous fungus with the ability to biocontrol other pathogenic fungus affecting crops through its mycoparasitism among other mechanisms of action. More recently, the capacity of using *Trichoderma* as a biocontrol agent for insect pests has been also considered. The use of *Trichoderma* spp. has allowed the commercial production of biocontrol agents for pest protection, growth enhancement of crops and the development of sustainable agriculture in different regions worldwide. In spite of its multiple applications, the metabolic and biochemical mechanisms of *Trichoderma* biocontrol capacity are only scarcely characterized.

Results:

To unveil the biochemical nature of Trichoderma's capacity to biocontrol different pests, we performed a genome-wide analysis to detect proteins evolving under positive selective pressure. Generally, proteins evolving under such patterns are linked to recent functional adaptations that could explain Trichoderma's biology. We used a recursive method to select different sets of orthologous proteins from OMA (https://omabrowser.org/oma/home/) for different species related to *Trichoderma harzianum*.10 different sets allowed us to collect homologues for ~60% of its proteome. For each set, we built DNA alignments using coding regions and the corresponding proteins alignments and this information was used to infer their evolutionary trees. Using PAML(http://abacus.gene.ucl.ac.uk/software/paml.html), we estimated the presence of positive selection at branch level (“branch models” using more than one dN/dS ratios in the tree) and at sequence level (“site models”, using 6 different models, M0, M1, M2, M3, M7 and M8). Models were compared using maximum likelihood ratio test and a chi squared statistical test.

Conclusions:

Our approach allowed us to identify 13 proteins with positive selection at the branch level possibly indicating functional adaptations of *T. harzianum* in reference to the other organisms. Estimating positive selection at sequence level, we also found that 115 proteins have residues evolving under positive selection also indicating functional adaptations of given positions. Our results will facilitate experimental design to uncover proteins endowing *T. harzianum* with its biocontrol capacity.

Poster Session: http://bit.ly/cab2c-2021-posters
Poster Aguer et al: https://youtu.be/p3Oh-7CrK8g

CoDNaS-Q: a high confidence first approach to explore the conformational diversity in homoligomeric proteins

Nahuel Escobedo[1], Ronaldo Romario Tunque Cahui[2], Gustavo Parisi[1], Alexander Monzon[3], Nicolás Palopoli[1]

[1] Departamento de Ciencia y Tecnología, Universidad Nacional de Quilmes - CONICET, Buenos Aires, Argentina

[2] Departamento de Ingeniería, Pontificia Universidad Católica del Perú, Lima, Perú

[3] Department of Biomedical Sciences, University of Padua, Padua, Italy

Background:

The concept of protein conformational diversity (CD) has been developing over the last 60 years promoting better understanding of the biological behavior of proteins. Under this paradigm, the native state of proteins is represented as a collection of conformers that together can describe the structural, binding and regulatory properties of a protein. CD has been extensively studied in proteins with a tertiary structure, however, at the quaternary structural level the CD characterization is scarce. We have developed CoDNaS-Q as the first database specifically aimed at studying CD at the quaternary level. CoDNaS-Q relies on experimentally determined CD, enabling the exploration of redundant collections of conformers with known homoligomeric structures in the Protein Data Bank (PDB).

Results:

We identified 18790 structures of sufficient quality that have a reported quaternary structure in the PDB. 3649 proteins were selected from sequence and structure comparisons, identifying the non-overlapping clusters of homoligomers, oligomeric states and structural superpositions. For each of these, CoDNaS-Q also lists information on features that could influence protein CD, such as its oligomeric state, length, pH, temperature, etc, as taken from the original structures.

Comparative studies of CD at the tertiary and quaternary levels allowed us to identify three groups of proteins with different dynamic behaviours, explained mainly by tertiary displacements, by rigid-body movements at the quaternary level or by a combination of both. Selected biological examples included in CoDNaS-Q showcase the relevance of CD to explain these behaviours. In fructose-1,6-bisphosphatase the union of the first substrate molecule in one dimer of the enzyme produces a conformational change in the other dimer, reducing the catalytic activity and affinity of its subunits, and thus providing an optimal therapeutic target for type 2 diabetes. Another interesting example is ferritins, a spherical shaped 24-mer protein which in the case of DpS (DNA protection during starvation) undergoes a conformational change at the tertiary level that interferes with the uptake and transformation of iron.

Conclusions:

The conformational diversity evidenced at the tertiary and quaternary levels suggests distinct dynamic characteristics in oligomeric proteins. The three groups of CD behaviours arising from CoDNaS-Q may be linked to alternative mechanisms in structure-function relationships. Comparisons between conformers of each protein, together with the description of their oligomeric states, ligands, experimental conditions and other data obtained from secondary information sources is available in our free online database CoDNaS-Q at http://ufq.unq.edu.ar/codnasq/.

Oral Presentation

In Silico analysis of genes for alkaloid metabolism in Lupinus mutabilis Sweet

Anais Mejia Solorzano[1], Raul Blas Sevillano[2], Christina Rivera-Romero[3], Jeny Perez-Huamanlazo[2], Julio, Solis-Sarmiento[1]

1.Facultad de Ciencias Biológicas. Universidad Nacional mayor de San Marcos
2.Instituto de Biotecnología. Universidad Nacional Agraria La Molina
3.Departamento Académico de Nutrición. Universidad Nacional Agraria La Molina

Background:
Lupinus mutabilis known as tarwi is the single domesticated lupin in the Andean region among four lupin species in the World . During last years, Lupin for human consumption has acquired relevance due to its demonstrated antidiabetic properties. Tarwi seeds is unique because they have the largest proportion of proteins among lupins and within legumes (up to 45% of seed dry weight) comparable to soybean; apart tarwi seeds have a high proportion of lipids(13.0-24.6 g/100 g dry weight) and significant amount of fibers, vitamins and minerals. The presence of toxic quinolizidine alkaloids (QAs) in the mature seeds (3%) is a negative trait that has impeded the . extensive cultivation and the utilization of tarwi seeds in comparison to other lupins. Currently, no genes associated with QA metabolism and transport has been reported for *L. mutabilis*. The main limitation for a wider use of tarwi in the Andean region is the lack of genomic tools, lack of a long term breeding program and the poor knowledge of the diversity of cultivars in the germplasm that may be adapted to different environment. Moreover, tarwi is very susceptible to pathogens such as *Colletotrichum* lupini, the causal agent of lupin anthracnose; source of resistance genes to this pathogen has not been identified in tarwi and also cultivars with low-alkaloid content are not known as compared to *L. angustifolius* and *L. luteus*.

Results:
By mining genomic data sets from the transcriptome and genome of *L. mutabilis* from our work and current public databases together with the genome of *L. angustigolius* we have started the identification of key enzymes involved in QA metabolism. Specifically, we have identified the genic fragment and their corresponding transcript of both lysine decarboxylase and c*adaverine* oxidase of *L. mutabilis* (LmLDC and LmCAO), orthologous of known genes of *L. angustifolius (*LaLDC and LaCAO) with 93 % and 94% of identity. LmLDC is an intronless gene (CDs of 1132 bp) while LmCAO has 12 exons (CDs of 5770 bp). We are conducting further work to integrate transcriptomic and genomic data from tarwi to the identification of other genes involved in QA metabolism.

Conclusions:
Comparative analysis of *L. mutabilis* and *L.angustifolius* allowed us to the identification and characterization of lysine decarboxylase and cadaverine oxidase genes from *L. mutabilis*, two key genes for the synthesis of quinolizidine alkaloids.

Poster Session: http://bit.ly/cab2c-2021-posters
Poster Mejia Solorzano et al: https://youtu.be/jp_0tMHq7qs

3D genome organization during human decidualization

Luciana Ant[1], Francois Le Dily[2], Silvana Panizzo[1], Griselda Vallejo[1], Francisco Pisciottano[1], Miguel Beato[2], Patricia Saragüeta[1].

1) Instituto de Biología y Medicina Experimental (IBYME-CONICET), Buenos Aires, Argentina.
2) Centro de Regulación Genómica (CRG), Barcelona, Spain.

Background:

With the advent of High throughput Chromosome Conformation Capture (Hi-C) analysis it has become possible to query the organization of the entire genome simultaneously at the sequence level trough the quantification of its interactions (Lieberman-Aiden et al., 2009). The resulting contact frequency maps are able to display self-associating domains formed by short-range interactions among contiguous segments of the genome and can be visualized as "triangles" present at the diagonal of Hi-C heatmaps. Classically, these contact domains are called compartments and Topologically Associating Domains (TADs).

Compartments are defined by Principal Component Analysis (PCA), normally using Hi-C data binned at 0.5-1.0 Mb resolution and, as a consequence, compartments are normally considered to be larger than 1 Mb in size. Compartments can contain sequences in an active (A) or silenced (B) transcriptional state and they interact with other compartments in the same state to give the plaid pattern observed in Hi-C heat-maps. On the other hand, neighbouring regions in separate compartmental domains interact less frequently and represent a compartmental switch or border. In this way, the compartmentalization of the genome creates both local compartmental domains and distant compartmental interactions.

In consequence, the chromatin organization can play an important role in transcriptional regulation, for example creating a microenvironment where genes regulated by a transcription factor aggregate together to coordinate their expression, or insulating an enhancer from its target gene to downregulate it.

Results:

In this work we explore the dynamics of the chromatin landscape and the transcriptional reprograming during decidualization of tHESC cells, a trans-differentiation process inherent of the human stromal endometrial cells. Decidualization was induced by exposure to cAMP, E2 and either progesterone or the synthetic progestin R5020 and Hi-C and RNAseq assays were performed at 60 minutes, 3 and 6 days later. Implementing only a few basic packages we were able to create custom Python scripts to extract the information from the raw data of RNAseq and Hi-C that provides some insight into the mechanisms of genome 3D organization that takes place in this process.

The PC1 values from the PCA of the Pearson correlation matrices were used to evaluate the global compartment reconfiguration, analysing correlation between different time points. The positive correlation (p-value<0.01) between B to A compartment changes and differential gene expression during human decidualization was also verified. Furthermore, heatmaps were directly generated from the aligned Hi-C reads using costumed Python scripts, without the need of any software or a lot of computing power.

Conclusions:

The analysis of the differential interaction heat-maps lead to the discover a specific region in chromosome 7, which comprises the Hox genes cluster, which showed dramatic changes in looping conformation, and also presented a correlation with changes in decidual specific genes expression after 6 days of treatment, indicating the increased interaction frequency between this cluster and the neighbouring gene promoters.

Oral Presentation

Design of specific primers for N gene of SARS-CoV-2 amplification by RT-PCR

Sarita Isabel Reyes, Maria del Milagro Said-Adamo and Hector Cristobal
Universidad Nacional de Salta
reyessarita.unsa@gmail.com, milagro.said@gmail.com, hacritobal@gmail.com

Abstract:
SARS CoV-2 causes severe respiratory syndrome, an etiological agent of the current COVID-19 pandemic. The World Health Organization approved detection of SARS-CoV-2 by RT-PCR screening of just one discriminatory target was considered sufficient. A mutation in the N gene (N-gene) of SARS-CoV-2 that adversely affects annealing of a commonly used RT-PCR primer was identified; epidemiologic evidence suggests the virus retains pathogenicity and competence for spread. This reinforces the importance of using multiple sequences of nucleotides for the specific SARS-CoV-2 primer design. The aim of this study was to design specific primers to amplify the complete nucleocapsid gene of SARS-CoV-2. Nucleotide sequences of N genes from different clade GR, G and GH circulating in Argentina and countries bordering were downloaded from the Global Initiative on Sharing All Influenza Data (GISAID). Multiple Sequence Alignment was carried out to identify the consensus regions on DNA-Man software using default settings. Different primer sets were designed using Primer Express® v.3.0.1 software. The primer sets obtained for bioinformatics analyses were tested in silico with a Basic local Alignment Search Tool (primer-BLAST) of public databases. The in vitro validation was carried out with Inactivated clinical samples from COVID-19 positive patients from Salta City. RNA was extracted with commercial kits. Retro-transcription (RT), End point PCR were performed using primers designed for N-gene amplification. The in silico results showed alignment with 100% of identity match and query coverage. The amplification length of the N gene was 1468 pb obtained with primer-BLAST software. High specificity was observed with primer designed for the virus target demonstrated by NCBI database analysis. Positive results for RT-qPCR reactions were obtained to amplify N and RNaseP genes from all samples. These results show correct sample processing, RNA extraction and amplification of SARS-CoV-2 genes by RT-qPCR. Nevertheless, in low concentration RNA SARS-CoV-2 positive samples, amplification of the complete N gene by RT followed by conventional PCR failed. This could be a consequence of poor integrity of RNA or RT-PCR enzyme efficiency. This study presents specific primers designed to amplify the complete N gene in COVID-19 patients from the province of Salta.

Poster Session: http://bit.ly/cab2c-2021-posters
Poster Reyes et al: https://youtu.be/QPMWqOg2UWQ

Exploring the genetic basis and evolution of Pantherine historical inter-specific hybridization

Clara Campos, Clementina Penna, David Bruque, Francisco Pisciottano and Patricia Saragüeta

[1] Instituto de Biología y Medicina Experimental - CABA, Provincia de Buenos Aires, Argentina [2] Unidad de Conocimiento Traslacional Hospitalaria Patagónica, Hospital de Alta Complejidad SAMIC - El Calafate, Provincia de Santa Cruz, Argentina

Background:

Genomic introgression analysis exposed a complex pattern of historical inter-specific hybridization in the Pantherinae lineage. This findings revealed unusually high rates of fertilization among pantherines, particularly between *Panthera leo* and *Panthera onca*, that would indicate an absence of inter-specific isolation barriers.

Zona Pellucida (ZP) and Juno (*IZUMO1R*) oocyte proteins play a key role in mammalian fertilization process, facilitating the interaction and fusion between the oocyte and the sperm respectively. Specifically ZP2 and ZP3 are directly involved in the interaction between gametes, ZP1 and ZP4 play the role of crosslinkers. Their interaction with sperm counterparts is essential for gamete interaction and the lack of recognition between them can constitute a pre-cigotic reproductive isolation mechanism. Hi-C long-range sequencing and chromosome-resolved genome assembly was used to reconstruct the intricate phylogeny of gamete interaction proteins to uncover the underlying molecular basis of inter-species crossing and evolution of the functional capacities of pantherines

Results:

In this work we explored the evolutionary patterns of ZP1, ZP2, ZP3, ZP4 and Juno proteins in the pantherines as well as in Feliformia and Caniformia suborders and the whole Carnivora order detecting adaptive changes, pairwise sequence identity and similarity analysis and evolutionary relevant sites based on curated multiple sequence alignments.

The conservation of all carnivora gamete interaction proteins was significantly higher (p-value < 0.05) in feliformes compared to caniformes except for the case of ZP4. ZP3 and ZP2 proteins displayed a similar pattern of evolution along their phylogenies showing adaptive changes in caniformes subtree but not among felid species. On the other hand, both ZP1 and ZP4 display a different evolutionary history, showing signatures of positive selection inside felids subtree. The fusion protein IZUMO1R did not show positive selection among all studied phylogenetic groups

Conclusions:

These results strengthen the idea that ZPs are proteins that provide a more specific-specificity isolation than IZUMO1R, which is specialized in fusion events. Altogether, our findings indicate that sperm-oocyte interaction and fusion proteins lack the degree of diversification necessary to fix a prezygotic reproductive isolation barrier in felidae.

Poster Session: http://bit.ly/cab2c-2021-posters
Poster Campos et al: https://youtu.be/YFnFn-RjrVo

Automatic annotation in GO based on Machine Learning in *Tetrahymena thermophila*

Costa,J[1]*; Bracalente,F.[1]*; Arabolaza, A. [1]; Gramajo, H[1].; Antonio D. Uttaro[1]; Spetale F.E[2]
[1]IBR-CONICET,Rosario,Argentina.

[2]Cifasis-CONICET,Rosario, Argentina.
costa@ibr-conicet.gov.ar
bracalente@ibr-conicet.gov.ar

* Both authors contributed equally

Background:

Tetrahymena thermophila is an unicellular ciliate that combines the complexity of cellular processes in eukaryotes with easiness in genetic manipulation and cultivation. In particular, the study of its sterol metabolism, that shifts from synthetizing the terpenoid alcohol “tetrahymanol” to assimilating and modifying sterols from its diet, when available, would help to understand cholesterol transport and homeostasis in higher eukaryotes. However, the mechanistic details of sterol uptake, intracellular transport and signaling systems in *T. thermophila* are still poorly known. The Gene Ontology (GO) terms associated with these functionalities are scarce in this microorganism, making it difficult to identify related proteins. Standard methods for the annotation of protein-coding genes based on sequence similarity, i.e., Blast2go, do not give good results. Therefore, automatic functional annotation methods based on machine learning (ML) rise as an alternative to standard methods. In this sense, this work aims to fill the gap in the annotation of *T. thermophila* proteins, emphasizing in GO terms of the sterol metabolism.

Results:

In this work, we predict the GO functionality of *T. thermophila* proteins using a novel graph-based ML package designed, FGGA, for the automatic annotation of proteins across the three GO subdomains. Proteins from *T. thermophila* with GO terms were collected from the UniProt database, and five GO terms involved in the sterol metabolism were enriched with proteins from related organisms.
The FGGA annotation algorithm assembles individual GO term predictions issued by binary SVM classifiers. Regarding the training of individual SVMs, a minimum of 50 positively annotated protein sequences was considered. In addition, to assemble conveniently balanced training datasets, positively annotated protein sequences were complemented with negative annotated protein counterparts using an inclusive separation policy. Concerning characterization methods of individual protein sequences in terms of a fixed number of input features, the measurement of 89 Pfam domains were considered. FGGA predictions were evaluated using a 20% test dataset and hierarchical Precision, Recall and F-score performance metrics. For the set of 535 GO-terms (BP-CC-MF) analyzed, we obtained 56%, 90% and 65% of the hierarchical Precision, Recall and F-score metrics, respectively. The validation process included three proteins (Q22MT, I7M195 and A4VD37) that showed a significant expression change in a previous RNA-Seq experiment under the presence of cholesterol. None has any annotation reported, being the algorithm able to classify them in sterol metabolism related GO-terms.

Conclusions:

The characterization of protein domain families has shown to be a valid approach for the classification of T. thermophila proteins.

Poster Session: http://bit.ly/cab2c-2021-posters
Poster Costa et al: https://youtu.be/jC7K8CN-_Yc

Towards a *lab-bench* Single-Cell Long-read transcriptomics (LRS scRNAseq)

Lavista Llanos S.[1], García Labari I.[1], Ezpeleta J.[1,2], Bulacio P.[1,2], Tapia E.[1,2]

1- CIFASIS-UNR/CONICET. Centro Internacional Franco-Argentino de Ciencias de la Información y de Sistemas
2-Facultad de Ciencias Exactas, Ingeniería y Agrimensura, UNR
lavistallanos@cifasis-conicet.gov.ar

Background:

Long-read sequencing (LRS) methods - Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT)- offer a number of advantages over short-read sequencing (SRS). LRS (multi-kilobase-scale) can mainly enhance the biological value of sequencing data from highly complex samples (e.g. single cell). However, LRS growth is still limited due to its unprecedented high rates of sequencing errors. This remains a huge hindrance for large-scale genomic applications like single-cell resolution studies that involve hundreds of thousands of samples.

Results:

We developed a robust method for large-scale LRS genomic applications based on our NS-watermark barcoding approach (Ezpeleta J. et al 2017). This new generation of NS-watermark DNA barcodes is inspired by *state-of-the-art* powerful coding methods used in digital communications and overcomes the challenging low sequencing accuracy (90%) of LRS that impairs a faithful barcode identification. We have validated our technology in experiments of 12, 24, and 96 multiplexed samples (MUX), and recently challenged its robustness in the order of thousands of barcodes. This faithful demultiplexing of thousand barcodes opens the door to LRS of single cells. Combining our new technology with a *lab-bench* single-cell barcoding method (Pacific Bioscience) we are assessing the individual transcriptome of ≈200.000 mammalian cancer cells as a proof-of-concept for our LRS scRNAseq technology. First experiments were performed in *mock* samples that helped us establish the molecular building blocks of this new protocol. In particular, we will apply this low-tech LRS scRNAseq assay to grasp molecular signatures that characterize acute- and chronic- bacterial infection conditions of human epithelial cells.

Conclusions:

Unlocking the true potential of LRS technology all while using standard lab techniques/equipment opens the door to high-capacity genome-scope studies of fundamental breakthrough potential (e.g., single-cell genomics).

Funding: PICT 2018-04480 and Argentag SAS - www.argentag.com

Poster Session: http://bit.ly/cab2c-2021-posters
Poster Lavista Llanos et al: https://youtu.be/-CE2kRTrk60

MACHINE LEARNING APPLICATION TO PREDICT THE DIAGNOSIS OF COVID19 BASED ON SYMPTOMATOLOGICAL PATTERNS

Albornoz, Germán[1], Mercedes Didier Garnham[1], Marcela Pilloff [1,2], Marina Mozgovoj [1,3], Maria José Dus Santos [1,4]

1 Universidad Nacional de Hurlingham
2 Departamento de Ciencia y Tecnología, Universidad Nacional de Quilmes
3 Instituto de Ciencia y Tecnología de Sistemas Alimentarios Sustentables (UEDD INTA CONICET), Argentina
4 Instituto Nacional de Tecnología Agropecuaria (INTA), Instituto de Virología e Innovaciones Tecnológicas (IVIT INTA CONICET), Argentina

Background:
A new type of coronavirus was identified in 2019 in the city of Wuhan, China. This new virus was called SARS-CoV-2 and is the etiological agent of the acute respiratory disease known as COVID-19, which spreaded around the world causing a pandemic. This disease affects patients differently, some present symptoms similar to a regular flu, such as fever, cough, odynophagia, myalgia, headache, among others. Other patients present severe symptoms, like bilateral pneumonia, severe acute respiratory syndrome, and septic shock.
One of the research fields that currently stands out in the fight against COVID-19 is Artificial Intelligence. Through the usage of computational models capable of learning to recognize patterns in a set of data, the symptomatic patterns of patients can be analyzed in order to determine which set of symptoms are representative of positive COVID-19 cases and thus, could be used to predict the diagnosis. From a clinical and epidemiological surveillance point of view, it is very important to know what specific symptoms are associated with the disease.
In this work, we selected 1500 symptomatic cases from health centers located in the districts of Hurlingham and Ituzaingó. Nasopharyngeal swabs from these cases had been processed at the COVID Unit belonging to the National University of Hurlingham (UNAHUR).
A dataset with the symptomatic pattern and their result for SARS-CoV-2 using RT-qPCR was created (n=750 positive patients and n=750 negative for COVID-19) and a Random Forest classification algorithm was used to analyze them.

Results:
The Random Forest analysis showed a 77% effectiveness and a 73% accuracy to discriminate between positive and negative cases. The sensitivity and the precision for the detection of positive cases was 93% and 70%, while for negative cases the values were 59% and 90% respectively. According to the model, the most relevant synthem to discriminate was the fever.

Conclusions:
The application of Machine Learning models, such as Random Forest, allows an accurate selection of positive cases and the prioritization of the patients to be tested. This would contribute with an efficient diagnostic procedure and the consequent optimization of available resources.

Oral Presentation