in person

24-25 Nov 2022

01001
00100
10100

XII CAB2C

Congreso Argentino de Bioinformática y Biología Computacional

Book of Abstracts

http://2022.a2b2c.org.ar

a2b2c

bioinformatica_ar

#XICAB2C

10100
10011

Electronic version of this book of abstracts:

XXXXXXXXXXXX

XII CAB2C

12vo Congreso Argentino de Bioinformática y Biología Computacional

Book of Abstracts

Asociación Argentina de Bioinformática y Biología Computacional - A2B2C

1ra Edición - 2022

Ciudad Autónoma de Buenos Aires

Libro Digital/PDF - Archivo digital: Descarga y Online

Este libro es una obra colectiva de los resúmenes enviados por sus autores y presentados en el 12vo Congreso Argentino de Bioinformática y Biología Computacional realizado los días 24 al 25 de noviembre de 2022 en la ciudad de Corrientes, Provincia de Corrientes, Argentina.

ISBN: XXXX-XXXX

Asociación Argentina de Bioinformática
y Biología Computacional

Presidente A2B2C
Dr. Sebastián Fernández Alberti

Vice Presidenta A2B2C
Dra. Lucía B. Chemes

Secretario
Dr. Nicolás Palopoli

Tesorera
Dra. Georgina Stegmayer

Vocales
Dra. Elin Teppa
Dr. Diego Bustos

Vocales Suplentes
Dra. Cristina Marino
Dr. Flavio Spetale

Comité Organizador
Dra. Lucía B. Chemes
Dr. Nicolás Palopoli
Dra. Elizabeth Tapia
Dra. Ana Julia Vélez Rueda
Dra. Juliana Glavina
Dr. Flavio E. Spetale
Dr. Tadeo Saldaño

Comité Científico
Dr. Máximo Rivarola
Dra. Georgina Stegmayer
Dr. Maximiliano Acevedo
Dr. Ariel Amadio
Dr. Sergio Samoluk
Dra. Lucía B. Chemes
Dr. Esteban Mocskos

01001
00100
10100

Our sponsors

Agencia I+D+i

CONICET

IDP f(un)

UNNE
Universidad Nacional del Nordeste

FaCENA - UNNE

ISCB REGIONAL Student GROUP Argentina

MBC
Mutual Bioquímica Correntina

Toyoko

ICCTI Instituto Chaqueño de Ciencia, Tecnología e Innovación

CORRIENTES

10011
01000

Table of contents

01001
00100
10100

Program

10100
10011
01000

Day 1: 24th November

09:00	Welcome Talk (Organizer Committee)
09:00 to 09:40	**KEYNOTE TALK. Alex Bateman (EBI-EMBL. UK)** *Structure Predictions Transform Protein Family Classification*
09:40 to 10:15	Break

TRACK: Proteomics, Protein Structure, and Function

10:15 to 10:45	**INVITED SPEAKER. Toby Gibson (EMBL. Germany)**
10:45 to 11:00	**Lightning talk. L. Bugnon, A. Edera, E. Fenoy, J. Raad, D. Milone and G. Stegmayer. (SINC-CONICET-UNL)** *Transfer learning to annotate (a part of) the Protein Universe*
11:00 to 11:15	**Lightning talk. L. Rodriguez Sawicki, G. Benítez, M.s Carletti, N. Palopoli, M.S. Fornasari and G. Parisi. (UNQ-CONICET)** *Conformational epistasis impairs AlphaFold structural predictions*
11:15 to 11:30	**Lightning talk. C. Lorenze, M. Safranchik, N. Garrone, J. Glavina and L. Chemes. (IIBIO-CONICET) (IIB-UNSAM)** *Identification and prioritization of SLiM-mediated interactions using Phage Display*
11:30 to 11:45	**Lightning talk. L.E. Fenoy and G. Stegmayer. (SINC-CONICET-UNL)** *Deep computational prediction of protein annotations combining sequence + structural learned embeddings*
11:45 to 13:00	Break - Lunch

TRACK: Comparative genomics, molecular diseases & evolution

13:00 to 14:00	**INVITED SPEAKER. Laura Kamenetzky (ib3- UBA, Argentina)** *Bioinformatic tools for whole genome analysis of helminth parasites*
13:30 to 13:45	**Lightning talk.J. Mac Donagh, D. Zea, G. Benítez, C.E. Guisande Donadio, J. Marchetti, N. Palopoli, M.S. Fornasari and G. Parisi. (UNQ-CONICET)** *Evolutionary rates in human amyloid proteins reveal their intrinsic metastability*

13:45 to 14:00	**Lightning talk.** **S. Prochetto, G. Stegmayer and R. Reinheimer. (IAL-CONICET-UNL) (SINC-CONICET-UNL)** *Self-organizing maps (SOM) based methodology reveals gene regulatory networks in plant evolution*
14:00 to 14:15	**Lightning talk.** **P.J. Salaberry and I.E. Schor. (IFIBYNE- UBA-CONICET) (DFBMC-FCEN-UBA)** *Tracking Key Regulators Of Metastasis In Triple-Negative Breast Cancer Through Combined Gene Regulatory Network Activity And Non-Coding Somatic Mutation Analysis*
14:15 to 14:30	**Lightning talk.** **M.S. Carletti, L. Rodriguez Sawicki, G. Benítez, M.S. Fornasari, N. Palopoli and G. Parisi. (UNQ-CONICET)** *Unveiling the origins of protein disorder using ancestral resurrection*
14:30 to 15:00	Coffee Break

TRACK: Genomics, Transcriptomics, and Metagenomics

15:00 to 15:30	**INVITED SPEAKER.** **Vinicius Maracaja-Coutinho (Brazil)**
15:30 to 15:45	**INVITED SPEAKER.** **Francesco Paolocci** *Uncovering more about the MBW complexes regulating proanthocyanidin biosynthesis in Lotus spp.*
15:45 to 16:00	**Lightning talk.** **R. Machado, S.n Moschen, G. Conti, J. Di Rienzo, S. González, L. Burdyn, E. Hopp and P. Fernández. (INTA)** *Discovery of new transcripts associated with HLB disease using a transcriptome-guided strategy*
16:00 to 16:15	**Lightning talk.** **L. Ant, F. LeDily, M. Beato and P. Saragüeta. (IBYME- Centro de Regulación Genómica (CRG))** *Multiomics approach in endometrial differentiation*
16:15 to 16:30	**Lightning talk.** **A. Baricalla, C.a Micaela, D. Del Valle, N.s Lavatti, C. Perez-Nieto, M. San Martin, C. Layana and R. Rivera-Pomar. (UNNOBA-CONICET)** *Unraveling the transcriptome of Andiperla morenensis*
16:30 to 16:45	**Lightning talk.** **E. Olszanowski, M.F. Ogara, A.S. Nacht, S.A. Rodríguez-Seguí, G.P. Vicent and A. Pecci. (IFIBYNE-UBA-CONICET) (FCEN-UBA)** *Lxr Activation Impairs Estradiol Dependent Proliferation In Human Breast Cancer Cells Through Downregulation Of Gene Expression Associated With Dna Replication And Cell Cycle Progression*

16:45 to 17:00	**Lightning talk. J.E. Kolomenski, C. Ballare, L. Dain and A.D. Nadra. (iB3-UBA-CONICET)** *Pipeline for NKX2-5 binding sites in the mouse and human genomes and possible interacting transcription factors*
17:00 to 17:15	Break
TRACK: Machine Learning	
17:15 to 17:45	**INVITED SPEAKER. Jessica Carballido (ICIC-CONICET, Argentina)** *Data analysis of gene expression*
17:45 to 19:30	**POSTER SESSION**
21:00	**SOCIAL DINNER**

Day 2: 25th November

09:00 to 09:40	**KEYNOTE TALK. Monica Pickholz (IFIBA-CONICET, Argentina)**

TRACK: Biological databases and Bioimaging

09:40 to 10:15	**INVITED SPEAKER. Luciano Abriata (EPFL)**
10:15 to 10:30	**Lightning talk. A.J. Velez Rueda, F.L. Bulgarelli, N. Palopoli and G.Parisi. (UNQ-CONICET)** *CaviDB: a database of cavities and their features in the structural and conformational space of proteins*
10:30 to 10:45	**Lightning talk. M. Bevilacqua, D. Clementel, A. Monzon, J. Lu, P. Arrias and S. Tosatto. (University of Padova)** *Improving Tandem Repeats Proteins annotation and classification in RepeatsDB*
10:45 to 11:00	**Lightning talk. M.L. Ingravidi, L. Dain, I. Ferder and L. Kamenetzky. (iB3-UBA-CONICET)** *In silico analysis of the expression levels of microRNAs predicted to target Fmr1: possible implications in Fragile X-Associated Disease*
11:00 to 11:15	**Lightning talk. L. Wiebke, J. Postat, J. Mandl, J. Textor and D.R. Parisi. (ITBA-CONICET)** *From the data to a minimal stochastic model for 1-dimensional T-cell dynamics*
11:15 to 13:00	Break - Lunch

TRACK: Big Data, Network and Machine Learning in Computational and Systems Biology

13:00 to 13:15	**Lightning talk. J. Raad, G. Stegmayer and D. Milone. (SINC-CONICET-UNL)** *RNABert: A novel RNA based embedding*
13:15 to 13:30	**Lightning talk. D. Clementel, S. Tosatto and F. Psomopoulos. (University of Padova)** *DOME Registry*
13:30 to 13:45	**Lightning talk. L.A. Bugnon, L. Di Persia, M. Gerard, A. Edera, J. Raad, S. Prochetto, E. Fenoy, G. Stegmayer and D. Milone. (SINC-CONICET-UNL)** *Improving the folding prediction of RNA with deep learning*

13:45 to 14:00	**Lightning talk. J. Gamboa. (CIDCA-CONICET)** *Unsupervised Machine Learning Approach To Quality Monitoring Of Strawberries During Drying*
14:00 to 14:30	Coffee Break

TRACK: IDPFun	
14:30 to 14:45	**Lightning talk. A.M. Monzon, A. Del Conte, D. Clementel, D. Piovesan and S. Tosatto. (Universita di Padova)** *MobiDB: intrinsically disordered proteins in 2022*
14:45 to 15:00	**Lightning talk. G. Erdos and Z. Dosztanyi. (MTA-ELTE)** *Disorder prediction based on estimated pairwise energies using deep learning methods*
15:00 to 15:15	**Lightning talk. M. Pajkos and Z. Dosztanyi. (Eötvös Loránd University)** *Identification of potential KLHL3 binding degrons based on machine learning approach*
15:15 to 15:30	**Lightning talk. J. Glavina, Z. Dosztányi and L.B. Chemes. (UNSAM-CONICET)** *Analysis of associations between motif-motif and motif-sequence features in intrinsically disordered regions*
15:30 to 15:45	Coffee Break
16:00 to 17:00	**DISPROT TRAINING SESSION**
17:00 to 17:30	**POSTERS AWARDS**
Closure of the Event	

01001
00100
10100

Keynotes and Invited Speakers

10100
10011
01000

Alex Bateman

Dr Alex Bateman joined EMBL-EBI as the Head of Protein Sequence Resources in 2012. He took over from Rolf Apweiler as the Principal Investigator for the UniProt grant (an international collaboration between EMBL-EBI, SIB and PIR) and has oversight for protein and ncRNA related databases at EMBL-EBI. Prior to that he managed the production of numerous popular biological databases such as Pfam and Rfam at the Wellcome Sanger Institute. He was formerly a member of, and Chairman of the ISB's Executive Committee. He was also Executive Editor for Bioinformatics and Editor of NAR's database issue.

Laura Kamenetzky

Laura Kamenetzky is a biologist (2001) and obtained a PhD in Molecular Biology (2007) from the School of Sciences (FCEN) at University of Buenos Aires (UBA). She did his post-doc at National Institute of Agricultural Technology (INTA). She is Independent CONICET Researcher at Institute of Institute of Biosciences, Biotechnology and Translational Biology- iB3-UBA) and develops her research in the area of genomics and bioinformatics studies of eukaryotes. She was the director of the Bioinformatics Node under the National System of High Performance Computing (SNCAD) until 2020. Nowadays she is Professor at School of Sciences FCEN-UBA. She coordinated research projects funded by Argentinian and European agencies. She organized several events and workshops in the areas of research in helminthes, bioinformatics and genomics. She is a reviewer for several high impact journals and national and international research agencies. She is author of 68 papers published in international journals (scopus h index 20) and 5 book chapters from national and international editorial.

Toby Gibson

Toby Gibson is at the European Molecular Biology Laboratory (EMBL) in Heidelberg, Germany since 1986. He studied Molecular Biology at Edinburgh University and did his PhD at the LMB, Cambridge, sequencing EBV. He is a computational biologist. He is a co-developer of the widely used Clustal series of multiple sequence alignment software. He oversees the development of ELM, the Eukaryotic Linear Motif resource (http://elm.eu.org/) devoted to protein sequence motifs involved in cell signalling and regulation.

Monica Pickholz

Licenciada en Física (FCEyN, 1994). Doctorado en la Universidade Estadual de Campinas (1999), seguida de estadías postdoctorales en el exterior. En marzo del 2009 se incorporó a la Facultad de Farmacia y Bioquímica a través del Programa RAICES, del MinCyT. En la actualidad es investigadora independiente de CONICET con lugar de trabajo en el Instituto de Física de Buenos Aires (IFIBA) en la Facultad de Ciencias Exactas y Naturales. La Dra. Pickholz posee una amplia experiencia con métodos de química cuántica y simulaciones de DM.

Jessica Carballido

Jessica Andrea Carballido works at the ICIC CONICET institute , and at the DCIC of the Universidad Nacional del Sur (Bahía Blanca, Argentina). She holds a PhD in Computer Science and is a university professor. She serves as an Adjunct Researcher at CONICET working in the area of Bioinformatics, category II incentives since 2018. Since 2000 she has been part of the Scientific Computing Research Laboratory, and at the beginning of 2018 she started an enriching and fruitful collaboration with members of the Cancer Biology Laboratory (INIBIBB) performing gene expression analysis. She is in charge of the direction of a Research Group Project accredited for incentives granted by the General Secretariat of Science and Technology of the UNS. She is an adjunct professor at the DCIC, in subjects in the area of programming, both undergraduate and graduate, with more than 20 years of teaching experience.

Vinicius Maracaja-Coutinho

Biologist from the Federal University of Paraiba (UFPB, Brazil) and PhD in Bioinformatics from the University of Sao Paulo (USP, Brazil), Vinicius is currently an Associate Professor at the University of Chile (Chile), where he leads the Laboratory of Integrative Bioinformatics and coordinates the Graduate Program in Bioinformatics and Computational Biology. In addition, he is also a permanent professor of the Master and PhD Programs on Bioinformatics at the Federal University of Rio Grande do Norte (UFRN, Brazil) and is the current president of the Chilean Bioinformatics Society. Vinicius is also the founder of the Chilean-Brazilian company Beagle Bioinformatics and has worked as a visiting researcher at the UFPB, at the Gurdon Institute from the University of Cambridge (UK) and at the IRD center in France; in addition to having worked as a consultant on missions for the International Atomic Energy Agency (IAEA), an organization linked to the United Nations.

Luciano Abriata

He is a scientist at the Laboratory for Biomolecular Modeling and the Protein Structure Core Facility in EPFL, Switzerland. Biotechnologist and PhD in Chemistry with experience in the wet and dry labs, he works on various structural biology/biophysics projects and guides the group's developments of novel human-computer interfaces for teaching, learning and working in chemistry and structural biology.

Francesco Paolocci

Francesco Paolocci completed his education as Agronomist at the University of Perugia. He is currently a Research Director at the Institute of Biosciences and Bioresources division of Perugia of the Italian National Research Council, where he has been coordinating teams studying genetics and genomics of plants. He taught Plant Biotechnology at the University of Urbino (Italy) for several years and he is currently a member of the board of the Ph.D. course in Biotechnology at the University of Perugia.

01001
00100
10100

Lightning talks

TRACK: Proteomics, Protein Structure, and Function

10100
10011
01000

POSTER# 4

Transfer learning to annotate (a part of) the Protein Universe

Leandro Bugnon, Alejandro Edera, Emilio Fenoy, Jonathan Raad, Diego H. Milone and Georgina Stegmayer

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

Background: The automatic annotation of proteins is still an unresolved problem. For example, as of August 2022, from 232,000,000 entries in UniProtKB only <1% of them are reviewed by expert curators. State-of-the-art annotation methods in Pfam, the protein family database, are based on hidden Markov models that predict family domain according to laborious hand-crafted sequence alignments. This approach has grown the Pfam annotations at a very low rate (<5% in the last 5 years). Alternative proposals based on deep learning models (DL) have appeared recently to accurately predict functional annotations for unaligned amino acid sequences. However, since many Pfam families contain just very few sequences, training such models with few examples is challenging.

Results: We propose to apply Transfer Learning for this task, that is, take advantage of pre-trained protein embeddings that integrate the information from millions of sequences in the complete UniProtKB. Nowadays there are several protein embeddings available and ready to use, such as ESM. It is based on BERT, a Transformer originally designed for Natural Language Processing, which is trained using the context to predict words. ESM makes an analogy between words and amino acids: it can learn meaningful encodings for each residue in a self-supervised way, by masking some of the residues in the sequence and trying to predict them as a pretext task. This way, the output sequence encodes context

residue information on each position. We obtained the ESM learned representation of the full domain data (17,929 families) from Pfam (1,339,083 seed sequences). Then, we used machine learning classifiers over the ESM embeddings in order to predict the domain for the test set (21,293 sequences), which had a low homology with the training set. We compared our approach with ProtCNN that is based on convolutional ResNets and achieved a 27.60% error rate (5,882 errors); while our method achieved a 20.88% error rate (< than 4,500 errors).

Conclusions: In this work we used cutting-edge transfer learning techniques to accurately predict protein domains. The results suggest that this approach presents unique predictive advantages and the potential to become a core component of future protein annotation tools.

POSTER #20

Conformational epistasis impairs AlphaFold structural predictions

Luciana Rodriguez Sawicki*, Guillermo Benitez*, Matias Carletti, Nicolas Palopoli, Maria Silvina Fornasari and Gustavo Parisi.

1 Departamento de Ciencia y Tecnología, Universidad Nacional de Quilmes. **2** INIBIOLP, Fac. de Cs. Médicas, UNLP-CONICET. *Equally contributed

Background: Protein structures have been massively predicted using homologous sequence information. These sequences are used to extract templates with structural information (such as in homology modeling) or sequence features which can be used to infer structural information (such as machine learning methods). AlphaFold2 (AF) is the last breakthrough to predict 3D models using machine learning approaches which reached an outstanding accuracy in the last quality evaluations. However, extant homologous information could not be enough to accurately predict protein structure. It has been shown that protein conformations depend on the evolutionary trajectory of substitutions in a process called conformational epistasis. Clear evidence of conformational epistasis was characterized in the evolutionary origin of the glucocorticoid (GR) and mineralocorticoid (MR) receptors specificity. From a MR- like ancestor (~440 million years ago (MY)), GR evolved high specificity for cortisol and diminished its affinity for mineralocorticoids between 420-440 MY. This process was characterized by the presence of four substitutions conserved in extant and ancestral GR. When these substitutions were introduced to extant MR to convert its phenotype to a GR-like activity, binding activity was abolished. However, when these substitutions were introduced in the MR-like ancestor, the obtained phenotype turned to be a GR ancestral state, with increased cortisol affinity and reduced aldosterone affinity. The observed difference between extant and ancestral behavior evidences conformational epistasis and the importance of evolutionary trajectories.

Results: In this work we explore how AF can reproduce conformations with epistatic effects. To that end, we obtained AF models for extant human GR and MR. Using 17 and 14 extant GR and MR homologous proteins with known structure, using principal component analysis we found that AF models for human GR failed to reproduce extant GR conformations. Also, other structural measures support these findings such as cortisol and binding sites distance comparisons. Interestingly, AF models for human MR showed to be almost indistinguishable from extant MR.

Conclusions: Our results showcase the importance of conformational epistasis to predict accurate 3D models using extant multiple alignments of homologous proteins. It is then possible that conformational epistasis could impair AF predictions since its use extant sequences without considering evolutionary trajectories.

POSTER #44

Identification and prioritization of SLiM-mediated interactions using Phage Display

Carla Lorenze[1,2]; Matías Safranchik[1,2]; Nicolás Garrone[1,2]; Juliana Glavina[1,2]; Lucía B. Chemes[1,2]

1 Instituto de Investigaciones Biotecnológicas, Universidad Nacional de San Martín - CONICET, Buenos Aires, Argentina. **2** Escuela de Bio y Nanotecnologías (EByN), Universidad Nacional de San Martín, Buenos Aires, Argentina.

Background: Short Linear Motifs (SLiMs) are short modular elements (~6 residues) located within disordered regions that mediate protein-protein interactions. The pocket protein family includes the Retinoblastoma (Rb), p107 and p130 proteins. Pocket proteins have a highly conserved domain that interacts with protein 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 (PD) assay that used the pocket domains as bait and a library of peptides from disordered regions of the human proteome (HD2) as preys.

Results: We identified overrepresented SLiMs in hits of the PD assay using SLiMFinder and MEME. We found an enrichment of the E2F and LxCxE motifs in the Rb screening while the p107 screening showed an enrichment only for the LxCxE motif and there was no enrichment of SLiMs for p130. We used IUPred to score peptides for disorder, and >70% of hits had an IUPred score of 0.4 or more, indicating these SLiMs are predicted to be highly accessible for interaction. Of the 29 known instances of the LxCxE and E2F SLiMs, 41 HD2 peptides had overlapping sequences with some of these instances. Among the hits, we recovered five of these instances for Rb, one for p107 and two for p130, a recall similar to published PD studies. The IntAct database revealed 156 interactors for Rb, p107 and p130, of which 111 were present in HD2. Among hits, we recovered 10% of the known interactors for each protein. These results suggest that it is possible to identify new proteins carrying functional motifs that interact with pocket proteins.

Conclusions: We conducted a proteome-wide screen of pocket protein binding SLiMs. We plan to extend the analysis using methods that allow the identification of peptides with a higher likelihood of being functional interactors based on identification by regular expressions, overlapping PFAM domains, and structure-based or AI-based scoring systems such as FoldX or AlphaFold2 to prioritize interactors for future *in vitro* validations. This approach will help expand the known pocket family interactome and understand its functions.

POSTER #7

Deep computational prediction of protein annotations combining sequence + structural learned embeddings.

Emilio Fenoy and Georgina Stegmayer.

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

Background: Improvements in experimental methods have generated rapid growth in the volume of protein data. In this scenario, the use of automatic methods has become critical to aid curation-based annotation. Many computational approaches to predict protein activities, properties, interactions, structure and functions have been proposed in the last few years. As a result of the increasing interest of the bioinformatician community regarding this topic, public benchmarks such as the Critical Assessment of Functional Annotation (CAFA) challenge were created, in which participants predict Gene Ontology (GO) terms for target proteins. However, the last results from the challenge report very low performance of the methods. We propose to increase the prediction rate by using representation learning. Nowadays, there are available many protein representation learning methods that calculate feature vectors for samples in a dataset. Such vectors, also known as embeddings, define a relatively low-dimensional space able to efficiently encode high-dimensional data. Recently, many protein representation methods have been developed, which integrate different types of protein information in supervised or unsupervised approaches, and provide embeddings based on sequence or protein structure . In this work, we propose a novel method to predict protein annotation by combining sequence and structural information encoded with state-of-the-art embeddings in a deep learning model.

Results: We used 9,500 Homo sapiens proteins from the CAFA3 to test the most recent sequence embedding methods and selected the best performing one, ESM-1b, a Transformer-based model trained over the full UniProtKB database. To complement the sequence embedding with structural information we predicted the tertiary structure of the proteins in the dataset using AlphaFold and used a novel structural embedding method called ESM-IF. This structure representation method was tested alone to predict GO annotations, obtaining state-of-the-art performance so far. Preliminary results on the combination of this method with the sequence embedding indicate that it is possible to boost their individual performance when used altogether.

Conclusions: The automatic annotation of proteins is a challenging task that requires new methods and strategies to be addressed. Here we state that combining sequence and structural information obtained from the current state-of-the-art methods such as AlphaFold, ESM-1b, and ESM-IF can improve the accuracy of the Gene Ontology annotation from a protein.

01001
00100
10100

Lightning talks

TRACK: Comparative genomics, molecular diseases & evolution

10100
10011
01000

POSTER #11

Evolutionary rates in human amyloid proteins reveal their intrinsic metastability

Diego Javier Zea*, Juan Mac Donagh*, Guillermo Benitez, Cristian Guisande Donadio, Julia Marchetti, Nicolas Palopoli, María Silvina Fornasari and Gustavo Parisi #

* Equally contributed

Proteins that are strongly driven towards aggregation in the form of amyloid fibrils are called amyloidogenic. In this work, we study the evolutionary rates of 81 human proteins for which an in vivo amyloid state is supported by experiment-based evidence. We found that amyloidogenic proteins evolve faster than the reference dataset (~16500 proteins from *Homo sapiens* with known orthologs on *Mus Musculus*), although they are highly expressed and abundant. In spite of the lack of significant differences in the evolutionary rates of secreted and amyloidogenic proteins, we found substantial differences in other features such as in their tendency to aggregate. Firstly, conformational diversity is higher in amyloidogenic proteins, also evidenced by the higher presence of disordered or highly flexible regions. The high conformational diversity could increase the chances of exposing amyloid-prone regions in slightly unfolded conformers driving the protein towards fibril formation. Secondly, amyloidogenic proteins are more expressed and abundant than secreted proteins. The impact of protein concentration and solubility on amyloid formation has been extensively studied. As it was described, to remain soluble, abundant proteins require the constant assistance of quality control mechanisms such as molecular chaperones. Furthermore, we found a positive correlation between RMSD100 (a measure of the degree of conformational diversity) and evolutionary rates which indicates that the higher the conformational diversity, the higher the evolutionary rates, possibly indicating a higher propensity to aggregate. Thirdly, a linear model combining an intrinsic protein parameter, as the RMSD100,with a cellular condition, as the supersaturation score, better explains the variation in the evolutionary rates observed in amyloidogenic proteins but not in secreted ones. Emerging evidence suggests that amyloidogenic proteins represent a “metastable subproteome” strongly driven towards the formation of amyloid fibrils. In the end, we showed that evolutionary rates reflect this particular behavior, showcasing the importance of metastability above other modulating factors. In the future, it could be interesting to evaluate protein metastability as a general modulating factor of evolutionary rates at the proteome level.

POSTER #12

Self-organizing maps (SOM) based methodology reveals gene regulatory networks in plant evolution

Prochetto S.[1,2], Stegmayer G.[2], Reinheimer R.[1]

1 Instituto de Agrobiotecnología del Litoral (UNL-CONICET), CCT-Santa Fe, Argentina. **2** Sinc(i), Research Institute for Signals, Systems and Computational Intelligence (UNL-CONICET), CCT-Santa Fe, Argentina. sprochetto@gmail.com; gstegmayer@sinc.unl.edu.ar; rreinheimer@ial.santafe-conicet.gov.ar

Background: Kranz syndrome is a set of leaf anatomical and functional characteristics of species using C4 photosynthesis. The current model for the evolution of C4 photosynthesis from a C3 ancestor proposes a series of gradual anatomical changes followed by biochemical adaptation of the C4 cycle enzymatic machinery. This transition occurred several times in various angiosperm lineages and is particularly interesting in grasses, where more than 1/3 of the C4 origins occurred. Despite decades of studies, the molecular mechanisms responsible for the evolution of photosynthesis remain unknown. In this work, leaf development traits from closely related C3 , C4 and intermediate species (Proto-Kranz, PK) were used together with gene expression data to study gene regulatory networks and identify potential drivers for the establishment of Kranz anatomy.

Results: Self-organizing maps (SOM) were developed to group features (genes and phenotypic traits) into clusters according to their expression along leaf development. Building several species-specific SOM and analyzing features displacement (changes in expression patterns) between species with different photosynthetic subtypes, allowed us to infer changes in gene regulatory networks and identify novel potential drivers for leaf development. The analysis between SOM showed general trends of specific gene displacements in the developmental network associated with photosynthesis, vascular development and transcription related processes. At the same time, a small subset of genes was observed to be displaced together with phenotypic traits, suggesting potential roles in the establishment of Kranz anatomy in grasses. This SOM based methodology applied to species-scaled transcriptomic and phenotypic data turned out to be a powerful tool to investigate the evolution of gene regulatory networks in leaf development.

Conclusions: The development of species-specific SOM for combining phenotypic traits and transcriptomics lead to the detection of displacements in gene regulatory networks between species with different photosynthesis subtypes, increasing the knowledge about the evolutionary path that leads to Kranz anatomy in grasses. Altogether, this method proved to be a useful tool for studying the evolution of gene regulatory networks and the detection of potential drivers of phenotypic differentiation.

POSTER #28

Tracking key regulators of metastasis in triple-negative breast cancer through combined gene regulatory network activity and non-coding somatic mutation analysis

Pedro J. Salaberry[1,2] 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: Whole-genome sequencing of tumor samples has 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 ability to identify genes and pathways associated with a more malignant phenotype. Then, we analyzed the frequency of potentially pathogenic regulatory mutations in their promoters in order to prioritize candidates through mutation recurrence.

Results: We combined a widely-used differential expression analysis method (DESeq2) and a breast cancer gene regulatory network (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, leading us to a list of 354 differentially active regulators. From this candidate set, we obtained an expanded network using the STRING database, and we delimited their active promoter regions in breast tissue using CAGE data from the FANTOM5 project. Finally, we retrieved reported TNBC mutations from COSMIC and PCAWG databases, and we used multiple scores to assess their pathogenic potential including FATHMM-MKL, CADD and DeepSEA. By means of a network propagation analysis, we found recurrently mutated sub-networks and functionally characterized them, identifying SWI/SNF as a protein complex with recurrent high-score non-coding mutations harbored in active promoters of its genes.

Conclusions: This work presents an effort to explore the potential use of regulatory mutations to understand cancer progression and identify key pathways driving malignant phenotypes. In the future, we intend to 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 ability.

POSTER #35

Unveiling the origins of protein disorder using ancestral resurrection

Matías S. Carletti[1] , Luciana Rodriguez Sawicki[1,2], Guillermo I. Benítez[1], M. Silvina Fornasari[1], Nicolás Palopoli[1], Gustavo D. Parisi[1].

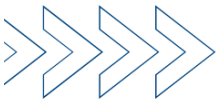
1 Departamento de Ciencia y Tecnología, CONICET, Universidad Nacional de Quilmes. **2** Instituto de Investigaciones Bioquímicas de La Plata. Fac. Cs. Médicas, UNLP-CONICET. gusparisi@gmail.com

Background: Conformational diversity is a key concept to understand protein biology. Protein movements can range from those involving residues, or those involving loops and secondary elements, to movements of entire domains. Proteins with extreme conformational diversity, or intrinsically disordered proteins (IDPs) are proteins showing a natively unfolded state characterized by the absence of secondary or tertiary elements. Here we study the origin of disordered regions (DR). Using resurrected proteins representing ancestral states of proteins (AncP) and their extant homologues, we can compare the occurrence of DR to analyze the origin, diversity and extension of DR in extant proteins. Our hypothesis is that DR in extant proteins may have arisen from diverse origins.

Results: Using the Revenant database, we detect actual homologous proteins to each AncP with known structure. We run BLASTP searches against the structures of 16 proteomes of model organisms from AlphaFold (AF) database. From a total of 18,230 structural alignments between the AncPs and their extant homologues, 23,912 DR distributed in 4,717 homologous AF structures were identified. From the comparison of each DR with the 55 AncPs, we can describe different scenarios: DR in extant proteins as arising from DR in AncP (Disorder-Disorder, DD), de novo appearance of DR (Order-Disorder, OD), or by inserting a DR in the AncP (Disorder by Insertion, DI) and a combination of the different mechanisms which we called mixed origin (MO). As a result of these comparisons we obtained the following distributions: 48.18% of regions of mixed origin where the major combination corresponds to OD_DI, 37.78% of OD, 8.95% DI, and the rest ~5% a combination of different mixed origins.

Conclusions: We conclude that DR has diverse and mixed origins. These results could contribute to explain the mechanistic origin of disorder and a putative classification of the different disordered regions in extant proteins.

01001
00100
10100



Lightning talks

TRACK: Genomics, Transcriptomics, and Metagenomics

10100
10011
01000

POSTER #33

Discovery of new transcripts associated with HLB disease using a transcriptome-guided strategy

Machado, R.[1], Moschen, S.N.[2], Conti, G.[3], Di Rienzo, J. A.[4], González, S.A.[3], Burdyn, L.[1], Hopp, H.E.[3], Fernández, P.[3]

1 EEA INTA Concordia. **2** EEA INTA Famaillá. **3** IABIMO-UEDD INTA-CONICET. **4** FCA, UNC. machado.rodrigo@inta.gob.ar

Multiple transcriptomic studies associated with citrus - Candidatus *Liberibacter spp.* interaction, causing HuangLongBing the most devastating disease of citrus, emerge annually. Nowadays, more than 25 RNA-Seq datasets were reported and the majority used a guided genomics strategy, since mandarin and orange genomes are available. However, transcriptome reconstruction has great advantages because it allows the complete transcripts assembly and quantification that represent multiple splice variants for each genetic locus. Transcriptome assembly can be performed by both genome-guided and genome-free. Several tools such as Cufflinks, iReckon, SLIDE and StringTie, incorporate existing annotations by adding them to the list of possible isoforms, which can be very useful for finding early biomarkers related to diseases. This approach reduce false discoveries that are common on reference-free assembly strategy. In the present PROCISUR funded study, multiple bioprojects (348468, 394061, 417324, 780217) were processed and individual transcriptomes were obtained. For that purpose, raw reads of each project were downloaded and adapters and low-quality reads were removed with fastp v0.23.2. The reads were aligned to the mandarin reference genome with STAR v2.7.8, and then transcript structures were predicted based on the reads aligned with StringTie v2.1.7. This unbiased approach allowed the comprehensive identification of all transcripts present in a sample, including annotated genes, novel isoforms of annotated genes, and novel genes. Then, the redundant transcript structures across the samples and the reference genome were combined using StringTie's merge function. The tool GFFcompare v0.11.2 was applied to evaluate the relationship from the merge transcriptomes and the reference annotation. GFFcompare evaluated the sensitivity (TP/(TP+FN)) and precision (TP/(TP+FP)) of each input transcript and computed at various levels (base, exon, intron chain, transcript, locus). In each case, around 45% novel exons, introns and loci were obtained, with good sensitivity (>90%) and precision (>60%) score. Novel isoforms were also visually analyzed in IGV. The discovery of new gene isoforms is useful for a better understanding of plant diseases, such as HLB, since these isoforms could be fulfilling various functions that were not being considered. However, each new isoforms must be biologically validated to be scientifically confirmed.

POSTER #39

Multiomics approach in endometrial differentiation

Luciana Ant[1], Francois LeDily[2], Miguel Beato[2] and P. Saragüeta[1]

1 IBYME. **2** Centro de Regulación Genómica (CRG)

Background: Chromatin topology is known to participate in gene expression changes associated with cell differentiation. To approach this issue in a hormone inducible system, we studied the conversion of stromal endometrial cells into secretory cells, a process known as decidualization, which is controlled mainly by progesterone.

Results: Hi-C and RNAseq analysis of tHESC cells were performed 1h, 3 and 6 days after hormone treatment. Then, custom python scripts were used to filter and merge the transcriptome and compartmentalization data from this experiment. Normalized contact matrices resulting from the Hi-C experiments were used to generate heatmaps depicting the contact frequency of different genome regions (5kb resolution). The transcriptional profiles were already modified at 1h in 96 genes, were TNFα pathway was significantly enriched (FDR<0,05). At 3 days, 42 of these early-regulated genes showed reorganization of chromatin topology in their regions. Also at 3 days, the cells exhibited a secretory profile and down-regulation of cell cycle genes (FDR<0,05) and a shift in chromatin state. Of the compartments that changed their state at day 3, 70,39 % did so from a closed to an open state . After 6 days, the cells continued expressing a secretory profile but with an increase in a subset of cell cycle genes. At this point, the changes in chromatin state occurred mainly from an open to a closed state.

Conclusions: The use of simple python scripts allowed us to obtain a general view of chromatin and gene expression state our system with minimal computational power. The questions that emerge from this analysis can be explored in detail using specialized software. It's interesting to know if the changes in compartment states at day 3 could be due to the transient cell cycle inhibition, since other gene expression pathways were not modified between day 3 and 6. However, other processes cannot be excluded. The genes up regulated at 1h might play a role setting off the differentiation process. Further studies at a higher resolution or single cell studies of chromatin accessibility could contribute to a better understanding of the relationship between genome topology and gene expression.

POSTER #63

Unraveling the transcriptome of Andiperla morenensis.

Agustin Baricalla[1,2], Micaela Coletta[1,2], Daiana del Valle[1,2], Nicolás Lavatti[1], Carola Perez[1], Maximiliano San Martin[1], Carla Layana[2,3], Rolando Rivera-Pomar[1,2]

1 UNNOBA. **2** CONICET. **3** Centro Regional de Estudios Genómicos. rrivera@unnoba.edu.ar

Insects are one of the most diverse groups within the metazoans, with a great variety of forms, developmental patterns, and habitats. Generally speaking, insects are active in a certain temperature range, which is predominantly between 20 and 30 °C. Above or below this, their activity ceases. Above or below this, their activity ceases. *Andiperla morenensis* is an insect that lives on the Perito Moreno glacier, i.e., its development surrounded by ice at temperatures below 4°C, making life impossible for any of its six-legged pairs. This work consists of the assembly of the transcriptome of Andiperla morenensis from a pool of insects, using TRINITY, in silico annotation using Trinotate, a pipeline that integrates BLASTx, BLASTp (both against Swissprot), HMMRScan against the Pfam base and the prediction of transmembrane domains and signal peptides. Manual curation of heat shock proteins, lipid metabolism, neuropeptides and their respective receptors was performed on these. We identified 5,206 transcripts related to mRNA metabolism and performed an analysis of the orthologs of the translation initiation factor 4E (eIF4E) and a prediction of their 3D structure. Finally, in order to understand the adaptation to low temperatures, the voltage-dependent sodium channel (VGSC) of the nervous system was studied. Related transcripts were searched, and the protein encoded by these transcripts shows the characteristics of the channels: transmembrane domains, voltage sensor and the amino acids of the inner (DEKA) and outer (EEQD) ring, as well as the inactivation domain (MFM). To corroborate the results, RT-PCR was performed on RNA obtained from larvae and the expression of this sodium channel and eIF4E in bacteria is being processed. These results provide the basis for the functional and structural study of the cold adaptation of *A. morenensis*.

POSTER #60

LXR activation impairs estradiol dependent proliferation in human breast cancer cells through downregulation of gene expression associated with dna replication and cell cycle progression

Evelyn Olszanowski[1,2], María Florencia Ogara[1], A. Silvina Nacht[3], Santiago Andrés Rodríguez-Seguí[1,4], Guillermo P. Vicent[5], Adali Pecci[1,2]

1 CONICET-Universidad de Buenos Aires, Instituto de Fisiología, Biología Molecular y Neurociencias (IFIBYNE), C1428EHA, Buenos Aires, Argentina. **2** Universidad de Buenos Aires, Facultad de Ciencias Exactas y Naturales, Departamento de Química Biológica, C1428EHA, Buenos Aires, Argentina. **3** Gene Regulation, Stem Cells and Cancer Program, Center for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Barcelona 08003, Spain. **4** Universidad de Buenos Aires, Facultad de Ciencias Exactas y Naturales, Departamento de Fisiología, Biología Molecular y Celular, C1428EHA, Buenos Aires, Argentina. **5** Molecular Biology Institute of Barcelona, Consejo Superior de Investigaciones Científicas (IBMB-CSIC), Barcelona, 08028, Spain.

Background: Liver X Receptors (LXRs) belong to the nuclear receptors superfamily of ligand activated transcription factors, whose endogenous agonists are the oxysterols. They play a key role in the regulation of the cholesterol homeostasis, induce the de novo synthesis of triacylglycerides, and counteract pro-inflammatory effects. LXRs are also known to compromise cell proliferation in several cancer models. However, their role in breast cancer (BC) has not been studied in depth and reports are, in fact, contradictory. Here we examined the potential involvement of LXRs in BC cells with special emphasis on their possible crosstalk with the Estrogen Receptor alpha (ERα). To address this objective, we performed colony formation (CFA) and propidium iodide staining assays in MCF-7 cells treated with or without Estradiol (E2) and the LXR agonist, GW3965. Our results showed that GW3965 impared the cell proliferation capacity induced by E2 (CFA: #colonies, Mean±SD: E2 208.7±25.7; E2+GW3965 131.3±23.7, n=3, padj<0.01, ANOVA).

Results: With the aim of understanding the molecular mechanisms underlying these functional effects, we performed a bulk RNA-seq experiment in duplicates. The differentially expressed genes were obtained using DEseq2, without using a cutoff for fold change (|log2FC|>0) to obtain not only the most differentially expressed genes, but also capture moderate effects. Between E2 and E2+GW3965 conditions we found several genes whose expression was affected by GW3965; which are widely enriched in terms associated to DNA replication, cell cycle, G1 to S transition, and Breast Cancer (padj<0.05) including genes such as PCNA, MCM4, CCND1, POLE3, TOP2A, BRCA1, BRCA2, RAD51, RET, E2F1, E2F2 as well as the ERα (ESR1). Interestingly, the presence of GW3965 increased the expression of the Glucocorticoid Receptor (GR) (NR3C1), which is consistent with a less proliferative phenotype observed in cells treated with this ligand.

Conclusions: Further experiments are necessary to address the mechanism underlying LXR function in BC, but our results point to a functional crosstalk with other steroid receptors, such as ERα and/or GR as putative mechanisms underlying LXR effects.

POSTER #38

Pipeline for NKX2-5 binding sites in the mouse and human genomes and possible interacting transcription factors

Jorge Emilio Kolomenski[1], Cecilia Ballare[2], Liliana Dain[1,3], Alejandro Daniel Nadra[1]

1 Departamento de Fisiología, Biología Molecular y Celular, Facultad de Ciencias Exactas y Naturales, Instituto de Biociencias, Biotecnología y Biomedicina - iB3, Universidad de Buenos Aires, Buenos Aires, Argentina. **2** Centre for Genomic Regulation - CRG, Barcelona, España. **3** Centro Nacional de Genética Médica, ANLIS, Buenos Aires, Argentina.

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. It's a transcription factor that is known to interact with others and form complexes and some of its genetic variants were found to be related with congenital heart disease. In order to better understand the binding sites of the NKX2-5 protein, we developed a pipeline to analyze experimental data and genomic sequences. Our workflow was aggregated into a standardized pipeline to identify binding sites of other transcription factors which could be interacting in complexes with the NKX2-5 protein.

Results: We developed a set of Python scripts to turn the standard outputs of ChIP-seq files into fasta files ready to be analyzed by MEME-ChIP, which returned potential binding sites. These analyses could reproduce the results seen in previous studies by other authors, which set a baseline for their functionality. All of this information was used to register NKX2-5 binding sites, nearby genes and co-occurrence with other transcription factor binding sites. Out of a total of 2610 ChIP-seq peaks in mice and 6871 in humans, we characterized 1110 and 2083 NKX2-5 binding sites close to genes in mice and humans, respectively. Genes were classified as "close" if they were within 50 kbp of the ChIP-seq peak. We also compiled the consensus binding sites for 51 proteins in mice and 169 in humans that could potentially interact with NKX2-5. Furthermore, we identified 588 ChIP-seq peaks not presenting any recognizable NKX2-5 binding sites in mice and 1984 in humans, suggesting indirect interaction.

Conclusions: This analysis allows for a guided study of binding sites in transcription factors in order to study their interactions as a network. The proteins for which binding sites were found in this study can be fed back into the pipeline in order to extend the studied network of interactions. We believe that this line of work can shed light into some of the interactions that control the formation of the early heart.

01001
00100
10100

Lightning talks

TRACK: Biological databases and Bioimaging

10100
10011
01000

POSTER #1

CaviDB: a database of cavities and their features in the structural and conformational space of proteins

Ana Julia Velez Rueda[1*], Franco Leonardo Bulgarelli[2], Nicolás Palopoli[1], Gustavo Parisi[1]

1 Departamento de Ciencia y Tecnología, Universidad Nacional de Quilmes - CONICET, Bernal, Buenos Aires, Argentina. **2** Mumuki.org

Proteins are the functional and structural units of cells. On their surface, proteins are sculpted into numerous concavities and bulges, offering unique microenvironments for ligand binding or catalysis. The dynamics of these cavities are fundamental for understanding protein function. The size and geometry of the cavities, as well as their accessibility, have proven useful in making predictions about protein-protein interactions, protein pharmacology, and binding specificity.

Here we present CaviDB (https://www.cavidb.org/), a novel database of cavities and their features in known protein structures, which integrates the results from commonly used software for cavities detection with protein features obtained from sequence, structure, and function analysis. We characterized a total number of 16,533,339 cavities, of which 62,0431 were predicted to be drug targets. CaviDB contains 276,432 different proteins with information on all their conformers. Each entry information is organized in sections, highlighting the general cavities descriptors, including the inter-cavities contacts, activated residues per cavity, and the information about druggable cavities, and the global protein descriptors. It also provides the ability to compare cavities and their properties from different conformational states of the protein.

Our database covers every protein structure available in the Protein Data Bank together with the predicted proteomes available for common health-relevant targets in EBI's Alphafold database. The data retrieved by the user can be downloaded in a format that is easy to parse and integrate with custom pipelines for protein analysis.

CaviDB aims to offer a comprehensive database for use not only in different aspects of drug design and discovery but also to better understand the basis of the protein structure-function relationship better. With its unique approach, CaviDB provides an essential resource for the wide community of bioinformaticians in particular and biologists in general.

POSTER #3

Improving Tandem Repeats Proteins annotation and classification in RepeatsDB

Martina Bevilacqua, Damiano Clementel, Alexander Monzon, Jiachen Lu, Paula Arrias and Silvio Tosatto

University of Padova

RepeatsDB is a database of structured Tandem Repeats Proteins (TRPs). Repeated units can be classified according to their shape, but their functional characterization and proper identification are still open questions. The goal of RepeatsDB is to identify the repeated units and their type on all available protein structures in the Protein Data Bank (PDB). An annotation is the association between the type of the repeat unit and its begin and end position on the protein structure. Annotations can be manually generated by a biocurator looking at the protein structure of interest (reviewed entries), or can be generated automatically through the RepeatsDB-lite predictor (unreviewed entries). RepeatsDB contains about 100,000 annotations over 7,000 PDB entries (~28% of total entries in PDB). In turn, such PDB entries are bound to just 1’500 UniProt entries. However, most of these are automatically generated. We implemented a distributed annotation framework, where anyone in the world should be able to annotate TRPs. However, we will still rely on a small group of trusted biocurators and reviewers which will grant that quality standards are met. This is helped by automatic statistical checks which would preemptively signal them any outlier or strange cases. Instead, non-trusted biocurators are rewarded through a badges-and-medals system based on gamification. The framework is scalable and can be easily applied to other classification tasks, outside the scope of RepeatsDB and TRPs classification.

POSTER #9

In silico analysis of the expression levels of microRNAs predicted to target Fmr1: possible implications in Fragile X-Associated Diseases

Marina Luz Ingravidi[1], Liliana Dain[1,2], Ianina Ferder[1], Laura Kamenetzky[1]

1 Instituto de Biociencias, Biotecnología y Biología translacional, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires. **2** Centro Nacional de Genética Médica, Administración Nacional de Laboratorios e Institutos de Salud, Buenos Aires, Argentina.

Background: The Fragile X Messenger Ribonucleoprotein 1 (FMR1) gene is located on the X chromosome and codes for the Synaptic Functional Regulator FMR1 Protein (FMRP). This gene is involved, by different molecular mechanisms, in 3 genetic disorders. Our group is interested in Fragile X Associated Primary Ovarian Insufficiency (FXPOI), a disorder that causes early menopause and sub-fertility/infertility, among others. Many studies show a correlation between the dysregulation of some microRNAs and the FMR1 transcript levels in different cell types of patients with Fragile X-associated diseases. Moreover, a microRNA cluster adjacent to Fmr1, Fx-mir, was recently described in mammals. Some of these microRNAs were shown to target FMR1 in mouse and human tissues. Our aim is to understand the regulation of Fmr1 by microRNAs in a model of follicular development in rat. In this work, we extended our previous research on the rat's Fmr1 regulators located in Fx-mir. To do so, we predicted all potential microRNAs that target Fmr1 and performed in silico analysis of the expression levels of microRNAs in the rat ovary.

Results: Using three specialized microRNA-target finding softwares (MIRANDA, DIANA, miRDB) we found 65 microRNAs predicted to target Fmr1. Five out of the top 30 were identified by the three softwares and 22 with at least two of them. We then searched for microRNA expression in the rat's ovary by analyzing publicly available small RNA-seq studies. We found 856 microRNAs expressed in the ovary. Among these, 57 were highly expressed (Reads Per Million, RPM>1000, high), 88 were moderately expressed (RPM>99, middle) and 711 had low expression (RPM<100, low). Within the forementioned 22 microRNAs that have Fmr1 as a potential target, 3 belong to the high group, 8 to the middle group, and 12 to the low group.

Conclusions: Even though most of the microRNAs predicted to target Fmr1 show a middle/low expression in the ovary, we found 3 microRNAs with a high predicted target score (rno-miR-148a-3p, rno-miR-92a-3p and rno-miR-182), that makes them good candidates to regulate Fmr1 in our model. We will experimentally validate these findings in our model of follicular development to understand their potential role in FXPOI.

POSTER #65

From the data to a minimal stochastic model for 1-dimensional T-cell dynamics

Wiebke L[1], Postat J[2], Mandl J[2], Textor J[3], Sultan S[3],Parisi D R[4]

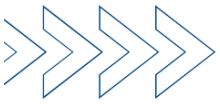
1 Centro de Agentes Físicos, Biológicos y Sociales. Instituto Tecnológico de Buenos Aires. 2 Faculty of Medicine, Department of Physiology, McGill University, Montreal, Canada. 3 Department of Tumor Immunology, Radboud Institute for Molecular Life Sciences, Nijmegen, Netherlands. 4 Centro de Agentes Físicos, Biológicos y Sociales. Instituto Tecnológico de Buenos Aires, CONICET

T-Cells dynamics is a relevant field of study because these cells need to move efficiently in extremely crowded environments like the lymph nodes. Understanding how it is accomplished and finding the maximum density where cells can still function may help in developing immunotherapy techniques. We aim to simulate high amounts of cells, so we need a computationally cheap model that represents key features of the system. We begin by studying the dynamics of mice T-cells on cell-wide microchannels (unidimensional system) using fluorescence microscopy. By properly understanding and representing simplest pairwise interactions, we expect to be able to also reproduce more complex emergent behaviours.

We develop a workflow from fluorescence microscopy images that allows extracting the cell trajectories used for analysing the experiments and developing our model. We include both existing tools for segmentation and tracking and develop our own scripts for further analysis. We describe phenomena like oscillations and group forming, with special attention to collisions. Finally, we develop an uni-dimensional agent based model that recreates the observed behaviour by following simple rules. Every agent moves straight in a given direction with a fixed speed until it either spontaneously reverts it, or it collides with another agent. In the second scenario, agents have to stop due to the collision, and until any or both of them reverts its velocity, they keep colliding. The probability of inverting its direction during a collision (alpha) is a variable parameter that can be used to tune the model with experimental data. We successfully reproduce the distribution of speeds, the distribution of train length and the duration of collisions.

The main strengths of the model are its scalability over the number of simulated agents and that it reproduces the features observed in the experimental data.

01001
00100
10100



Lightning talks

TRACK: Big Data, Network and Machine Learning in Computational and Systems Biology

10100
10011
01000

POSTER #13

RNABert: A novel RNA based embedding

J. Raad, G. Stegmayer and D.H. Milone

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

Background: Embedding methods are being actively proposed by applying deep learning to biomolecular information. Biological data, such as sequences, require an embedding operation that encodes its unstructured data into a numerical vector. In the deep learning field, embeddings pre-trained with a large set of unlabelled data has shown to be effective for many downstream supervised-learning tasks, even when a smaller size of labelled data is available. Obtaining better embeddings enhances the quality of downstream analyses, such as function or secondary structure prediction. Recently, DNA sequences have been effectively embedded using deep representation learning, with techniques developed in the field of natural language processing. As a consequence, word embedding techniques have been applied to nucleotides for DNA sequences, obtaining DNABert.

Results: In this work we used the pre-trained DNABert for the effective embedding of RNA bases to provide semantically rich representations and better predict RNA secondary structures. The DNAbert pre-trained model with 6-mer tokenization was used and a fine-tune adjustment was performed in two stages: a first self-supervised stage with 125,000 RNA sequences from Homo Sapiens obtained from Rfam, where a token of the sequence is masked and then the model has to predict it from its context. Next, the last linear layer of the model was eliminated, and a new linear layer was added where a second supervised fine-tune was performed with 400,000 RNA sequences to predict the 4,096 families of Rfam. Finally, the obtained embedding was combined with a deep Residual Network (ResNet) to predict the secondary structure of an RNA from its sequence alone. Assessed in cross-validation, our RNAbert+ResNet model achieved a notable increase in F1 compared to DNAbert+ResNet.

Conclusions: In this work we use transfer learning to reuse valuable information from the DNA domain in the modeling of RNA sequences, obtaining an embedding that can be used for better RNA representation and downstream tasks. Furthermore, the results obtained in the prediction task suggest that applying transfer learning to RNA sequence modeling can help improve the prediction of secondary structures from a few labeled examples.

POSTER #26

DOME Registry

Damiano Clementel[1], Silvio Tosatto[1] and Fotis Psomopoulos[2]

1 University of Padova. **2** Greece Center for Research and Technology Hellas.

Thanks to high-throughput technologies, large amounts of biological data are being generated and made accessible to researchers. Machine learning has come into the spotlight as a very useful approach for understanding biological data. This led to the corresponding growth of machine learning publications. However, only a few of them are experimentally validated. Therefore, there is the need of establishing well defined community standards, in order to increase openness and reproducibility of experiments. Guidelines or recommendations on how to appropriately construct machine learning algorithms can help to ensure correct results and predictions. On the ML community side, there is demand for a cohesive and combined set of recommendations with respect to data, the optimization techniques, the final model, and evaluation protocols as a whole (DOME). DOME recommendations referenced in this work specifically tackle supervised learning for biological applications. The DOME Registry is a database which provides a curated set of annotations for papers about machine learning, following the DOME recommendations. described above. The database is provided as a website, which allows users to search among the 158 manually annotated papers and to visualize information as described by the guidelines. It also provides a statistics page. Among various statistics, annotation goodness is evaluated by means of the DOME score. Moreover, authenticated users can access an insertion form, where DOME compliant information can be added to the database. Through the development of the DOME registry, we hope to provide a tool for easing the search of information in the machine learning findings within the biological community. Moreover, the registry can be used by authors to promote their work and enhance its quality by filling the input form which enforces the information to follow the DOME recommendations. Doing this, we aim at establishing the DOME Registry as a reference for high quality information for machine learning within the biological community and beyond.

POSTER #18

Improving the folding prediction of RNA with deep learning

L.A. Bugnon, L. Di Persia, M. Gerard, A. Edera, J. Raad, S. Prochetto, E. Fenoy, G. Stegmayer and D.H. Milone.

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

Background: The function of noncoding RNAs (ncRNAs) is relevant for numerous biological processes and largely depends on their secondary structure, which determines interactions with partner molecules. However, the determination of RNA structures is a very costly process, which cannot be scaled up efficiently, limiting our ability to functionally characterize such molecules. Computational methods are promising for the prediction of RNA structures, which is speeding up the discovery of function and action mechanisms. Since classical tools strongly rely on hand-crafted thermodynamic features, they show limited capacity for modeling the wide structural diversity of RNAs, including pseudoknots, non-canonical linkages and long sequences. Recently, machine learning techniques have demonstrated being able to capture thermodynamic features in an automated manner.

Results: We have compared several recent methods for secondary structure prediction, including classical and newer ones based on deep learning, using 3023 yeast sequences and a novel benchmark of well-characterized long ncRNA from different species. Classical methods can guarantee a 50%-70% prediction accuracy in all the range of RNA in the dataset. However, their weakness is that their results have stagnated in the last 20 years. Despite being relatively new, machine learning based methods have achieved similar predictive performance to the classical ones for long ncRNAs. We propose a new architecture for secondary structure prediction based on machine learning. The network first encodes information about each nucleotide and its neighbors from which nucleotide-to-nucleotide information is then used to predict the connection matrix. This allows representing any type of connection including pseudoknots. This architecture was evaluated on a curated dataset obtained from RNAs with experimentally verified structures from public repositories. Preliminary results show that our proposal can reach a high performance.

Conclusions: In this work, we identify current limitations of ncRNA folding prediction tools, especially with long ncRNA, and propose new methods to improve them. Using a curated dataset and a benchmark to test model generalization, promising results have been achieved. Computational limitations will be considered in order to process large sequences (>1000nt).

POSTER #58

Unsupervised machine learning approach to quality monitoring of strawberries during drying

Gamboa-Santos, J*

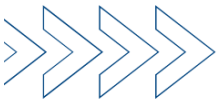
CIDCA (CONICET-CCT y Universidad Nacional de La Plata), 47 y 116. La Plata, 1900, Argentina, j.gamboa@conicet.gov.ar

Background: This work aims to analyse the potential of unsupervised machine learning (ML) models to be applied during food quality monitoring of drying operations. In this sense, a microwave (MW) assisted drying (1.2 W/g, 100 min) of strawberry was carried out to validate a machine vision system composed of a digital camera and a 6-modules system using image embedding, programmed in Python. The system collected information about several morphological and colour features (area, enclosing rectangular area, height, width, radius, brightness and saturation retentions). Feature extraction tasks were performed from 490 individual images of fresh (FR) strawberry samples during MW drying processing. Final dataset accounted the information of 19 features (9310 samples), among them, drying kinetics parameters (moisture losses at each drying time condition) were also computed. With the aim to identify drying time thresholds of quality changes a k-Means clustering model was performed combining the selected features.

Results: After applying a dimensionality reduction to 3 principal components, FR samples subjected to MW drying were naturally clustered among two (k=2) groups, with high clustering metrics performance. By analysing 16 different clustering configurations (with k values of 2 or 3) the better results were obtained for two drying time categories (k=2) at configuration 4 (C4, ARI: 0.957, AMI: 0.915), suggesting a quality threshold at drying times above 60 min.

Conclusions: The results presented here showed the potential of unsupervised classification methods coupled with image embedding as an attractive and economical alternative system to define quality thresholds in order to in-line monitoring of quality changes during drying.

01001
00100
10100



Lightning talks

TRACK: IDPFun

10100
10011
01000

POSTER #14

MobiDB: intrinsically disordered proteins in 2022

Alexander M. Monzon, Alessio Del Conte, Damiano Clementel, Daminao Piovesan and Silvio Tosatto

Universita di Padova

The MobiDB database (URL: https://mobidb.org/) provides predictions and annotations for intrinsically disordered proteins as long as binding regions and experimentally-derived contact annotations. Latest update of MobiDB (version 4) includes novel types of annotations and an improved update process. The new website has been re-designed, with a new user interface, a more effective search engine and advanced API for programmatic access. The new database schema gives more flexibility for the users, as well as simplifying the maintenance and updates. In addition, the new entry page provides more visualization tools including customizable feature viewer and a sequence viewer that is responsive to the actions made on the feature viewer. Moreover, with the new AlphaFold2 revolution, we were able to integrate a new structure viewer for the majorities of the entries in MobiDB. This will permit the user to interact with the predicted structure, coloring the annotated regions of the entry present in the database. The structure viewer will also switch to the PDB structures when they are available for a particular region. MobiDB v4 annotates the binding modes of disordered proteins, whether they undergo disorder-to-order transitions or remain disordered in the bound state. In addition, disordered regions undergoing liquid-liquid phase separation or post-translational modifications are defined. The integrated information is presented in a simplified interface, which enables faster searches and allows large customized datasets to be downloaded in TSV, Fasta or JSON formats. An alternative advanced interface allows users to drill deeper into features of interest. A new statistics page provides information at database and proteome levels. The new MobiDB version presents state-of-the-art knowledge on disordered proteins, disordered/ordered regions are assigned based on AlphaFold pLDDT score and improves data accessibility for both computational and experimental users.

POSTER #48

Disorder prediction based on estimated pairwise energies using deep learning methods

Gabor Erdos and Zsuzsanna Dosztanyi

MTA-ELTE Momentum Bioinformatics Research Group

Many proteins contain intrinsically disordered regions (IDRs), functional polypeptide segments that do not form a well-defined structure in physiological conditions. Disorder prediction methods, which can discriminate ordered and disordered regions from the amino acid sequence, have contributed significantly to our current understanding of the distinct properties of intrinsically disordered proteins by enabling the characterization of individual examples as well as large-scale analyses of these protein regions. One popular method, IUPred provides a robust prediction of protein disorder based on an energy estimation approach that captures the fundamental difference between the biophysical properties of ordered and disordered regions. Derivative methods of IUPred such as ANCHOR allow for the identification of functional disordered regions. These energy estimation based methods offer a very fast and reliable way for the identification of disorder properties. In this work I extended the original ideas behind IUPred and ANCHOR and added a novel deep learning approach to enhance the capabilities of these prediction methods. Our current show a significant improvement compared to our old results which are further confirmed by the preliminary results of the second CAID proposal.

POSTER #54

Identification of potential KLHL3 binding degrons based on machine learning approach

Mátyás Pajkos and Zsuzsanna Dosztanyi

Eötvös Loránd University, Faculty of Science, Institute of Biology

The ubiquitin-proteasome system has an essential role in the regulation of many diverse biological processes (cell cycle progression, DNA repair, genome integrity, etc.) via targeting specific substrates for proteasomal degradation. Abnormalities of this system are associated with various diseases, such as Alzheimer, Parkinson, Huntington and several types of cancer. A key step of the proteasomal degradation is the substrate recognition and its ubiquitination, which is carried out by E3 ligases. In most cases, the substrate recognition is mediated by short linear motifs, called degrons. Although the human proteome contains more than 600 E3 ligases, the substrate binding mechanism and the specific degron motifs have been characterized for only a handful of them. The main aim of this project is to identify novel binding motifs for KLHL3 protein through a bioinformatic pipeline. KLHL3 is a member of the KLHL family that forms the substrate recognition subunit of Cullin-RING type E3 ligases. For the KLHL proteins, the degron binding is carried out by the extremely conserved Kelch domain. Evolutionary analyses show that the KLHL proteins are abundant in humans (more than 40 identified paralogs), which indicates their biological significance. For KLHL3, the only well studied substrates are the human WNK isoforms. It has been shown that missense mutations in the WNK4 degron or in the motif binding region of the KLHL3 Kelch domain causes Gordon's hypertension syndrome. Similarly to the WNK4 substrate, additional KLHL3 binding degrons can be involved in disease development. We established a motif identification bioinformatics pipeline to find novel degrons for the human KLHL3 protein using machine learning based approach. In this search procedure, we exploit several types of sequence features, such as sequence disorder, secondary structure and disorder binding prediction, and combined them using a random forest algorithm. Using our tool, we could search the complete human proteome database for proteins that contain the binding degron, identifying proteins that are likely true binding partners. The identified motifs can help us to gain further insights into the biological role of KLHL3 protein and its role in disease development.

POSTER #36

Analysis of associations between motif-motif and motif-sequence features in intrinsically disordered regions

Juliana Glavina[1,2], Zsuzsanna Dosztányi[3], Lucía B. Chemes[1,2]

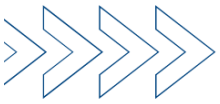
1 Instituto de Investigaciones Biotecnológicas, Universidad Nacional de San Martín - CONICET, Buenos Aires, Argentina. **2** Escuela de Bio y Nanotecnologías (EByN), Universidad Nacional de San Martín, Buenos Aires, Argentina. **3** Department of Biochemistry, Institute of Biology, ELTE Eötvös Loránd University, Budapest, Hungary

Background: The function of an intrinsically disordered region (IDR) can be determined by short linear motifs (SLiMs) which mediate protein-protein interactions. These interactions are encoded by 3-5 residues that belong to the motif core, but the regions flanking the SLiM can also modulate the strength of the interaction. In addition, there is growing evidence from several case examples that IDRs evolve as a whole, and that coevolution and functional coupling of SLiMs are key components for functional encoding within IDRs. Here, we analyzed the co-occurrence of experimentally verified SLiMs with other sequence elements to determine unknown functional couplings that may help to elucidate the function of IDRs.

Results: We collected over 1000 protein sequences harboring experimentally described SLiMs annotated in the ELM database. Using statistical tests, we identified ~100 significant SLiM-SLiM and SLiM-PFAM domain associations. In the first case, 20% of the associations were between pairs of SLiMs annotated in ELMdb. For example, the PipBox1 SLiM which mediates binding to PCNA is found in proteins involved in DNA replication, repair, and cell cycle control. PipBox1 showed associations with other SLiMs that modulate binding to PCNA and with the MutS and XPG PFAM domains found in proteins related to DNA repair and translesion synthesis. We used bootstrapping to determine the correlation between SLiMs and significantly enriched sequence features in the N- and C-terminal 12 residues flanking the SLiMs. We found ~50 associations of SLiMs with enrichment in proline content and net charge among others. PipBox1 was associated with an increase in net positive charge in the flanking region preceding the SLiM, in agreement with experimental evidence that the N-terminal flank (140-RKRR-143) of the p21 PipBox1 SLiM enhances binding to PCNA.

Conclusions: A proteome-wide analysis allowed us to identify multiple associations of annotated SLiMs with predicted SLiMs, PFAM domains, and multiple sequence features. Some of these associations are known to be functionally relevant, suggesting that the remaining associations identify yet unknown functional couplings between SLiMs and other functional elements within IDRs. Our approach improves our knowledge of SLiM functions and opens up avenues for the identification of novel functions within IDRs.

01001
00100
10100



Poster presentations

10100
10011
01000

POSTER #2

Assessing the similarities between alternative promoters in human genome to understand the impact of promoter architecture on transcription

Martin Iungman[1], Ignacio E. Schor[1,2]

1 IFIBYNE UBA-CONICET. **2** DFBMC FCEN-UBA.

Background: Several studies over the last decades have identified the core promoter's architecture as an important determinant of a gene's transcriptional properties, including regulatory selectiveness, genetic robustness and noise. In this context, the existence of genes with multiple promoters that can be regulated between cell types or conditions, poses an interesting model to study the importance of core promoter features in transcription.

Results: Using two stages of an in vitro neuronal differentiation process in human cells as a model, we studied the covariance between transcription start site (TSS) clusters usage and transcriptional noise genome-wide. We used available data from Cap Analysis of Gene Expression (CAGE) and single-cell RNA-seq to measure TSS relative activity and transcriptional noise respectively. We used generalized linear models to test for a global association between these two variables, failing to observe a significant influence of promoter usage on transcriptional noise. In light of the results, and considering the known effects of promoter sequence on this feature, we considered the possibility that, for genes with more than one promoter, the evolution could have favored the similarity between the architecture of these promoters. In order to explore this hypothesis, we started analyzing the distribution of promoters' architectural features for alternative promoters of the same gene, and comparing this with random promoter associations. As a first step, using promoters derived from FANTOM5 CAGE annotations, we observed a significantly higher number of genes that present uniformity in the presence of TATA-box among alternative promoters than what is expected by chance.

Conclusions: In addition, this analysis presented several challenges that we are currently working to deal with. For example, since some of the CAGE-derived TSS clusters are too close, they could be parts of the same promoter. Therefore, we are developing a pipeline to uncover pairs of promoters that functionally act as a unit, using the variance in the activity of these TSS clusters across biological samples. We believe that how architectural properties of promoters impact gene function and how genes and promoters are born and evolve are two important and connected questions, and that functional genomics can contribute valuable information to answer them.

POSTER #5

Bioinformatic tools to study the function of FABP5 in lung adenocarcinoma

Costa ML[1], Mancini EM[2], Scaglia N[1]

1 INIBIOLP, National University of La Plata/National Council of Scientific and Technical Research of Argentina. **2** Senior Bioinformatics Manager MultiplAI 3-7 Temple Avenue, Suite 140, Temple Chambers, London, United Kingdom, EC4Y 0DA

Background: Lung cancer is responsible for the highest proportion of cancer-associated deaths in Argentina. Previous work from our laboratory showed that Fatty Acid Binding Protein 5 (FABP5) has substantial effects on lipid metabolism, cell proliferation in lung adenocarcinoma (LUAD) cells in vitro, as well as tumor growth in vivo. The aim of this work is to study the role of FABP5 in gene expression regulation in LUAD and to improve our knowledge of the molecular mechanisms of this pathology in order to find novel therapeutic targets.

Results: To assess the role of FABP5 as a transcriptional regulator, we performed transcriptome sequencing (RNA-Seq) of LUAD cells under pharmacological FABP5 inhibition. Using an established bioinformatic pipeline for Differential Gene Expression (DGE), we detected 638 DGE under the 2 conditions. We confirmed our results with the analysis of RNA-Seq data of tumor samples from LUAD patients from The Cancer Genome Atlas database. After we filtered the samples according to the purity of the tumor and divided them according to the expression levels of FABP5. We confirmed that patients with high levels of FABP5 in tumors have worse overall survival. We obtained 725 DEG when comparing samples with high vs low levels of FABP5. A functional enrichment analysis using GSEA showed numerous pathways associated with the cell cycle affected by FABP5 expression, both in LUAD cells and tumors. To investigate possible transcription factors (TF) involved in the control of gene expression, we performed TF binding domain discovery assays using the XSTREME tool of The MEMESuite, followed by a domain enrichment analysis. Remarkably, we found several potential TF involved in transcriptional regulation by FABP5.

Conclusions: Our results extend the knowledge about the role of FABP5 in the regulation of transcription in LUAD and support our previous findings. The implications of FABP5 in the overall survival of patients with LUAD highlights the need to elucidate the mechanisms through which it exerts its action and its biological consequences.

POSTER #6

Enzymatic rational design guided by structural bioinformatics: Enhancing the affinity of PP1 to microcystin-LR with saturated mutagenesis followed by docking

Ezequiel Alba-Posse[1,2], Carlos David Bruque[3,4], Javier Gasulla[1,2], Alejandro Daniel Nadra[1,2]

1 Universidad de Buenos Aires. **2** Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET). **3** Unidad de Conocimiento Traslacional Hospitalaria Patagónica, **4** Hospital de Alta Complejidad El Calafate SAMIC.

Background: The group is working on the development of an enzymatic colorimetric biosensor based on the inhibition of the Ser/Thr phosphatase PP1. The objective is to improve the biodetection of the cyanobacterial toxin microcystin-LR (MC-LR) in water samples.

Results: From the 6270 possible variants of PP1, we filtered 129 candidates that incremented both affinity and stability according to a FoldX Stability and Autodock Vina rigid docking compared to wild-type. Those 129 were reduced to 19 after full-flexible active site docking stage. From them, 3 variants were chosen for in vitro evaluation (p.Q249Y; p.D197F; p.S129W). Site-directed mutagenesis were performed on PP1 plasmids with the Q5 Hot Start kit (NEB). E.coli BL21 were transformed with the mutated plasmids and each recombinant mutant protein was expressed and purified. They were evaluated in stability (by fluorescence spectroscopy), activity (pNpp dephosphorylation measured at 405 nm), and affinity to MC-LR (percentage of inhibition compared to wt). Although 2 of the 3 proteins (p.S129W; p.Q249Y) showed less affinity or significantly less activity, one of the 3 proteins (p.D197F) not only showed significantly more affinity to MC-LR than wild type, but also presented significantly more activity in solution. The proposed mechanism is due to the increment in hydrophobicity (ASP to PHE) of the hydrophobic pocket of PP1, where MC-LR “Adda” aromatic residue buries.

Conclusions: In order to obtain PP1 variants that perform better in an activity assay used for MC-LR biodetection, we conducted a saturated-mutagenesis bioinformatic screening. 3 variants were chosen for in vitro evaluation. One of the proteins (p.D197F) not only showed significantly more affinity to MC-LR than wild type, but also presented significantly more activity in solution.

POSTER #8

Molecular docking: computational methods to predict protein-carbohydrate interactions

Jorge Octavio, Rafael Betanzos, Carlos Pablo Modenutti

CONICET - UBA, IQUIBICEN

Background: Molecular docking is a bioinformatic method that allows computationally predicting the binding site and the most favorable position of interaction between a ligand and a given target (usually a protein). AutoDock4 (AD4) is the most popular docking program in the world. It's primarily designed to align rigid drug-like fragments into the binding sites of macromolecules but, when applied to flexible carbohydrate molecules, frequently displays poor performance due to their low affinity, their hydrophilic nature and ligand conformational flexibility. The overall objective of this work is to develop a computational method capable of efficiently estimating the ligand pose at the carbohydrate binding site, using crystallographic structures as reference. For this purpose, a dataset of different crystallographic structures of protein-carbohydrate complexes was used. For each complex, the ligand molecule was removed from the target model and then docked back (redocking) applying different modifications in the protocol in order to test the effects of each parameter.

Results: The parameters affecting docking accuracy are: ligand flexibility and complexity, bias selection and application, and receptor structure and conformation. Selecting these parameters correctly allows to achieve the expected results.

Conclusions: The best way to improve carbohydrate docking efficiency and accuracy using AD4 is achieved by setting correctly the empirical values of ligands glycosidic linkages' torsion angles, and by applying biases based on solvent structure at the binding site. Even so, obtained results are highly conditioned by the protein structure. Reducing the error range that this entails is the next challenge within carbohydrate docking, which aims to improve the reliability, quality and reproducibility of computational predictions. The above would enhance the use of comparative or homology-based models when no crystal structure is available, which is vital for understanding key biological recognition processes and for glycomimetic drugs development.

POSTER #10

In silico evaluation of the interaction of peptide ligands with bevacizumab by structural analysis and molecular docking

Juan F Orlowski[1], Gabriela R Barredo-Vacchelli[1], Silvana L Giudicessi[1], Silvia A Camperi[1]

1 Universidad de Buenos Aires, Facultad de Farmacia y Bioquímica, Cátedra de Biotecnología. Instituto NANOBIOTEC UBA-CONICET, Junín 956, (1113) CABA, Argentina.

Background: Bevacizumab, a monoclonal antibody used for cancer treatment, binds vascular endothelial growth factor (VEGF) preventing angiogenesis. Bevacizumab must be highly purified to be intravenously administered by an expensive process based in affinity chromatography (AC) with protein A immobilized. Alternatively, AC with short peptides as ligands reduce the cost of the process and are of higher physically and chemically stability. Recently, synthetic peptides have been evaluated for bevacizumab AC purification. The best results were obtained with a peptide corresponding to the 85-PHQGQHIG-92 VEGF. However, the screening process was very expensive and laborious. The aim of this work was to probe if in silico studies could predict peptide interaction and reduce the time and cost of the screening assay.

Results: The 3D peptide structure was modeled with PEP-FOLD3. The molecular docking was executed with AutoDock Vina software. Five models were found through directed molecular docking simulation with good performance in molecular coupling when evaluated for their binding energy. The conformations obtained against the same fragment within the native crystallized structure were compared. These models' root mean square deviation (RMSD) values were of about 2 Å which implies that they were positioned in a similar way to the crystallized structure. The comparison of the crystallographic structure of the complex VEGF/bevacizumab against the models, conducted in LigPlot+, allowed to identify the key amino acids and the type of interactions involved. The H-bond interactions between the peptide and Gly33 and Thr53 of bevacizumab were preserved in both the simulated models and the native structure together with the hydrophobic interactions.

Conclusions: Here it was demonstrated that molecular docking analysis is a useful tool to evaluate in silico ligand-target interactions. The correlation of the bioinformatic results with the experiments may save time and resources in ligand design to optimize protein purification as well as drug developing.

POSTER #15

Human apolipoprotein B100 protein prediction and modeling of the interaction with the alkaloid methyl cytisine by molecular docking.

Antonella M. Bucci[1], Maria de los A. Frias[1], Ana E. Ledesma[1,2]

1 Centro de Investigación en Biofísica Aplicada y Alimentos (CIBAAL-UNSE- CONICET), Universidad Nacional de Santiago del Estero, RN 9, km 1125, (4206) Santiago del Estero, Argentina. **2** Universidad Nacional de Santiago del Estero, Facultad de Ciencias Exactas y Tecnologías, Departamento Académico de Química. Av. Belgrano Sur 1912, 4200, Santiago del Estero, Argentina.

Background: Low-density lipoprotein (LDL) particles are the main carriers of cholesterol in human plasma. The organization of the particle, which is composed of apolar lipids and a monolayer of phospholipids stabilized by apolipoprotein B100 (Apo-B100), as recognized by LDL receptors, is highly complex and still unknown. ApoB 100 is an extremely large protein (4563 amino acids) and very little is known about its structure, different experimental approaches to resolve the tertiary structure of LDL, have the inconvent of a complex particle organization. Reports of isoquinoline-type alkaloids, such as berberine (benzyltetrahidroxyquinoline), have demonstrated their potential as an antihyperlipidemic agent by reducing serum levels of cholesterol, triglycerides, and LDL cholesterol due to its antioxidant activity by inhibition of LDL lipid peroxidation.

Results: Five domains were also reported in terms of trypsin accessible peptides. the sequence data reported in UNIPROT for each domain, structure prediction corresponding to lipoprotein domains was modeled using the I-TASSER, and AlphaFold2 server. The structure of the ligand N-Methyl-cytisine, (NMC), alkaloid was previously obtained in our group. The 3D structure of apolipoprotein B100 was generated sequentially (eeach domain), the set of domains was obtained by calculating protein-protein coupling using the HADDOCK2.2 web server. From this result, the structure of the complete protein was validated using the graphic server Ramachandran observed that 85% of residues falling in the highly favorable zones. The secondary structure predicted by the YASARA program.

Conclusions: The binding site between the N-methyl cytisine with each domain of apolipoprotein B100 and were studied by molecular docking using the AutoDock 4.2 program, the free binding energy of the I domain presents the lowest value independently of the rest of the protein, which indicates that this site is the most probably site for interaction, this current The residues involved in the interaction were characterized, finding polar and hydrophobic residues as responsible for the stabilization of the complex with the NMC alkaloid. More studies are being carried out to evaluate its effect in the oxidative process of apolipoprotein B100.

POSTER #16

Gene function prediction in five model eukaryotes exclusively based on gene relative location through machine learning

Flavio Pazos Obregón[1,2], Pablo Soto[3], Rafael Cantera[3], Diego Silvera[3], Patricio Yankilevich[4], Gustavo Guerberoff[5]

1 Uruguay Instituto de Investigaciones Biológicas Clemente Estable. 2 Instituto Pasteur de Montevideo. 3 Instituto de Investigaciones Biológicas Clemente Estable. 4 Instituto de Investigaciones en Biomedicina de Buenos Alres. 5 Facultad de Ingeniería, Universidad de la República.

The function of most genes is unknown. The best results in automated function prediction are obtained with machine learning-based methods that combine multiple data sources, typically sequence-derived features, protein structure, and interaction data. Even though there is ample evidence showing that a gene's function is not independent of its location, the few available examples of gene function prediction based on gene location rely on sequence identity between genes of different organisms and are thus subjected to the limitations of the relationship between sequence and function. Here we predict thousands of gene functions in five model eukaryotes (*Saccharomyces cerevisiae*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Mus musculus*, and *Homo sapiens*) using machine learning models exclusively trained with features derived from the location of genes in the genomes to which they belong. Our aim was not to obtain the best-performing method for automated function prediction but to explore the extent to which a gene's location can predict its function in eukaryotes. We found that our models outperform BLAST when predicting terms from Biological Process and Cellular Component Ontologies, showing that, at least in some cases, gene location alone can be more useful than sequence to infer gene function.

POSTER #17

Weighted Gene Co-Expression Network Analysis using sunflower public repositories ranks and identifies candidate genes in defense response to fungus

Andres Ribone[1,2], Sergio Gonzalez[1], Maximo Rivarola[1,2]

1 Instituto de Agrobiotecnología y Biología Molecular (IABIMO), INTA-CONICET, Hurlingham, Argentina. **2** CONICET, Buenos Aires, Argentina

Background: In the past years, GWAS and QTL mapping efforts have been widely used in sunflower to identify hundreds of gene candidates for fungal resistance. These lists can be quite lengthy in large QTL regions, thus needing a way of prioritizing loci. This problem is more evident when functional annotations are sparse. Insight may still be uncovered using public genomic and transcriptomic repositories; a more systemic approach to connect existing data is becoming more and more vital to gain knowledge to continue plant breeding.

Results: To tackle the above mentioned task, we utilized all the public accessible transcriptomic data from sunflower tissues to construct several gene co-expression networks. Our initial reference network was constructed from 686 RNAseq samples of healthy green tissue. Its quality was evaluated via its capacity to connect genes with shared GO terms, and the relation between gene connectivity and rates of molecular evolution. As expected, more connected genes have lower dN/dS than peripheral genes. Moreover, the mean adjacency between genes with shared GO terms was much higher than expected by chance. In consequence, Guilt-By-Association (GBA) GO term prediction showed an overall high performance. We identified clusters of genes enriched in defense functions and focused on genes in such clusters not annotated as such. Moreover, we were interested in four clusters which presented the enriched GO term “defense response to fungus”, of which one had 33 candidate loci identified in previous QTL mapping studies. This same cluster had 8 out of the 37 WRKY transcription factors found in our network, and a total of 24 uncharacterized/unknown function genes. In particular, one uncharacterized hub gene has an homolog in *Arabidopsis thaliana* which has been described to be affected by several stress conditions. Moreover, the GBA method predicted it as “defense response” associated at an F1 score of 0,82.

Conclusions: We present our work and validate our methodology to existing knowledge and show its capability to identify/rank new candidates for crop breeding programs. Our future goal is to construct, integrate and contrast different biological networks and provide additional functional annotation to candidate genes in defense response to fungus.

POSTER #19

protAGOnist: an innovative NLS/NES prediction tool

Camila Engler[1,2], Belen Moro[2], Andrea Martin[2], Antonela Lavatelli2[·], Luciano Abriata[3], Nicolás Bologna[2]

1 Faculty of Engineering, National University of Entre Ríos (UNER), Argentina. **2** Centre for Research in Agricultural Genomics (CRAG), CSIC-IRTA-UAB-UB, Bellaterra, Barcelona 08193, Spain. **3** École Polytechnique Fédérale de Lausanne (EPFL), Switzerland.

Background: RNA silencing controls gene expression via 19-36 nucleotide small RNAs to regulate adaptive responses to stress, preserve genomic integrity by controlling transposon activity, conduct general innate immune response to viruses, and play a key role during development, among others. To fulfil their biological function, small RNAs are loaded into ARGONAUTE (AGO) proteins that recognize and regulate their target genes. Depending on the organism and the small RNA pathway, ARGONAUTEs have nuclear and/or cytoplasmic localization. The lack of reliable prediction tools is a significant obstacle in determining the subcellular localization of ARGONAUTE proteins, which limits our knowledge of the specific subcellular functions of ARGONAUTE proteins and our understanding of the mechanism of intracellular movement of small RNA. Standard sequence-based predictions often suggest a high number of false positives nuclear localization and/or nuclear export signals. Combining these sequence predictions with the biophysical properties of amino acids, their evolutionary conservation, and molecular modelling, we have developed a computational tool that improves the scoring of true nuclear localization signals and nuclear export signals, reducing false positive rates. This tool not only takes into account the primary amino acid sequence as a standard prediction program but also the three-dimensional structure, exposure, flexibility and region of each subcellular localization signal to efficiently reduce the false-positive signals that are very common in standard prediction programs.

Results: The output includes a CSV file with the analyzed data and the label true or false for each putative signal, and a graph to visualize the position and signal. We already applied protAGOnist to more than 20 ARGONAUTE proteins in *Arabidopsis thaliana*, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Mus musculus* and human. While standard nuclear localization and nuclear export signals prediction programs detect more than 250 signals. Applying our tool we were able to improve significantly the NLS/NES prediction reducing highly the number of false positives. We also were able already to validate several signals obtained by protAGOnist in several ARGONAUTE proteins from different eukaryotic organisms

Conclusions: The results show that the computational method developed has the ability to reduce the number of putative signals, which would facilitate the functional analysis of the proteins under study.

POSTER #21

Dynamics of SARS-CoV-2 during the first year of the COVID-19 pandemic in Northwestern Argentina

Zambrana Montaño R.[1, 2], Culasso A.[1,2], Fernández F.[3], Marquez N.[3], Debat H.[3], Salmerón M.B. [4], Zamora A.M.[4], Ruíz de Huidobro G.[4], Costas D.E.[4], Alabarse G.[4], Charre M.A.[5], Fridman A.D.[5], Mamani C.[5], Vaca F.[5], Maza Diaz C.[5], Raskovsky V.[6], Lavaque E.[6], Lesser V.[6], Cajal P.[6], Agüero F. [6], Calvente C.[6], Torres C.[*1, 2], Viegas M.[*7]

1 Universidad de Buenos Aires, Facultad de Farmacia y Bioquímica, Instituto de Investigaciones en Bacteriología y Virología Molecular (IBaViM), Buenos Aires, Argentina. **2** Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Buenos Aires, Argentina. **3** Instituto de Patología Vegetal, Centro de Investigaciones Agropecuarias, Instituto Nacional de Tecnología Agropecuaria (IPAVE-CIAP-INTA), Córdoba, Argentina. **4** Laboratorio de Salud Pública, San Miguel de Tucumán, Tucumán, Argentina. 5 Laboratorio Central de Salud Pública, San Salvador de Jujuy, Jujuy, Argentina. **6** Laboratorio de Virus Respiratorios y Neurovirosis. Hospital Señor del Milagro, Salta capital, Salta, Argentina. **7** Laboratorio de Virología, Hospital de Niños Dr. Ricardo Gutiérrez, CABA, Argentina. *****Shared last authorship.

Background: Diversity and evolution of SARS-CoV-2 is reflected by variants and lineages. While variants of concern (VOCs) and variants of interest (VOIs) have been defined by the WHO, hundreds of lineages have been described under the Pango system. This study aimed to analyze the evolutionary dynamics of the main SARS-CoV-2 lineages circulating during the first and the early second wave of the COVID-19 pandemic in Northwestern Argentina.

Results: Phylogenetics and phylodynamic analyses were performed on SARS-CoV-2 whole-genome sequences from the provinces of Jujuy (n=93), Salta (n=89) and Tucumán (n=78) from individuals diagnosed with COVID-19 from March 2020 to May 2021 in Argentina. In the first wave (March 2020 - February 2021), eight lineages were identified from the 108 SARS-CoV-2 genomes in the phylogenetic analyses (Maximum Likelihood method, IQ-TREE v2.1 software): B.1.499 (76.9%), N.5 (10.2%), B.1.1.274 (3.7%), B.1.1.348 (3.7%) and others in minority. For the predominant lineage B.1.499, there were monophyletic groups between two provinces (Jujuy and Salta), which suggests common transmission chains, but not with Tucuman. In the early second wave (March - May 2021) an increase in the lineage diversity (12 lineages forthe 152 genomes) and faster replacement dynamics was observed: while lineage N.5 predominated during the beginning of the second wave (April 2021), numerous introductions of lineages P.1 (VOC Gamma) and C.37 (VOI Lambda) were observed and increased later (May 2021). Phylodynamic analyses (BEAST v1.10.4 software) for the two main Argentinean lineages that circulated during the studied period revealed that the rate of evolution of lineage N.5 (7.9×10^{-4} s/s/y) was a ~40% faster than that of lineage B.1.499 (5.6×10 -4 s/s/y), although both are in the same order of magnitude than rates for other non-VOC lineages. Besides, the demographic reconstruction of the lineage B.1.499 showed an increase in the effective number of infections that matched the reported cases, evidencing its role in driving the first wave in the region.

Conclusions: The main first-wave lineages in the NWA showed particular evolutionary patterns, being displaced at the early second wave mainly by the introduction of the VOC Gamma and the VOI Lambda.

POSTER #22

Genome- Wide characterization of Dof gene family in peanut (Arachis hypogaea)

Samoluk SS[1,2]; Seijo JG[1,2]

1 Instituto de Botánica del Nordeste (CONICET- UNNE), Corrientes, Argentina. **2** Facultad de Ciencias Exactas y Naturales y Agrimensura (UNNE), Corrientes, Argentina. samocarp31@gmail.com

Background: The Dof gene family encodes a group of plant-specific transcription factors. Studies of these genes in some agronomic important crops have revealed that play a key role in plant growth, development, and response to biotic/ abiotic stresses. The recent publication of peanut genome and their wild diploid progenitors provides a great opportunity to perform genomic approaches for a better understanding of the regulation of many important traits, a crucial step for crop improvement. In the present study, we report a genome-wide analysis of Dof genes in A. hypogaea (AhyDof), including phylogenetic inferences, gene structures, chromosomal locations, gene duplications events, and the analysis of expression patterns.

Results: A total of 61 full- length genes (AhyDofs) across 20 chromosomal pseudomolecules of peanut genome were detected. Although the deduced AhyDof proteins were variable in length, molecular weight and isoelectric point, they showed a highly conserved zinc finger structure, characteristic of this family. Phylogenetic analysis revealed the existence of different groups of sequences with particular distribution patterns of protein domains, as well as exonic and intronic structures. Genome- scale analysis of synteny indicated that segmental duplication events may have played an important role in the evolution of the AhyDof genes. Analysis of cis- acting elements in the promoter region revealed the presence of motifs associated with response to hormones, light and stress conditions. The expression levels of these genes were variable in the different tissues and developmental stages of the plant.

Conclusions: This study provides insights into the evolution and functional diversification of AhyDofs as a basis for future research in cultivated peanut.

POSTER #23

Biog5: A bioinformatic system for the analysis of the Human Papillomavirus

Ernesto Rafael Pérez[1], Sofía Erdozain[2], Leonardo Gómez Chávez[1], Germán Conti[1], Emilio Angelina[1], Nélida Peruchena[1]

1 Laboratorio de Estructura Molecular y Propiedades (LEMyP), Instituto de Química Básica y Aplicada del Nordeste Argentino, (IQUIBA-NEA). **2** Facultad de Ciencias Exactas y Naturales-Universidad de Buenos Aires.

Background: Human papillomavirus (HPV) is classified according to its oncogenic potential into different types of risk which are low, high and undetermined. Currently, there are more than 100 different types but just the most prevalent and high-risk ones (nine of them) are currently covered by available vaccines. In addition, there are several less common strains that can still cause cancer and be resistant to vaccines. Biog5 is a tool that integrates multiple computational biology and bioinformatics tools to find correlations between all requested HPV strains in sequence analysis and structural data of L1, L2, E1, E2 and E7 proteins. This allows us to assist in diagnostic and outbreak investigations to address the virulence and likelihood of vaccine resistance of the less prevalent strains.

Results: Biog5 uses public databases and alignment algorithms, Hidden Markov Models, phylogenetic trees, and conservation analysis to predict whether a selected HPV strain can cause a serious health condition and whether available vaccines might be effective. Component testing on HPV strains of well-known phenotypes yields correct classification to risk groups and high similarity between strains of the same group.

Conclusions: The trend and accessibility of sequencing technologies hint at their potential use within our healthcare systems. With this in mind, Biog5 was conceived as a tool that could be useful, for example, in epidemiological surveillance, that is, for the planning, implementation and evaluation of public health programs based on the phenotypes of local HPV strains. In addition, it could also be used in patients with HPV, for the prognosis of the outcome of the disease.

POSTER #24

Using Model-Driven Approach to model and simulate Tissue Engineering construct with Multiagent Systems.

Angelini José[1,3,4,5], Molas Giménez Tomás[2], Pérez Ángeles[3], Moreyra Jesús[1], Muñoz Justo[1], Romagnoli Javier[1], Fernández Peterson Aníbal[1], Di Paolo José[5] and Gutiérrez María de los Milagros[6]

1 Cátedra de Comportamiento Físico de Biomateriales, Carrera de Bioingeniería. Facultad de Ingeniería de la Universidad Nacional de Entre Ríos. Ruta 11 km 10, Oro Verde, Entre Ríos. **2** Laboratorio de Cibernética, Dpto. Bioingeniería Facultad de Ingeniería, Universidad Nacional de Entre Ríos. Ruta 11 km 10, Oro Verde, Entre Ríos. **3** Universidad Tecnológica Nacional Facultad Regional Santa Fe (UTN-FRSF). Lavaise 610, Santa Fe. **4** Cátedra de Materiales Dentales I. Carrera de Odontología. FCS. Universidad Adventista del Plata. Villa Libertador San Martín. Entre Ríos. **5** Grupo de Biomecánica Computacional. Facultad de Ingeniería de la Universidad Nacional de Entre Ríos. Ruta 11 km 10, Oro Verde, Entre Ríos. **6** CIDISI Universidad Tecnológica Nacional Facultad Regional Santa Fe (UTN-FRSF). Lavaise 610, Santa Fe. jose.angelini@uner.edu.ar

Background: The influence of activated genes by Growth Factors (GF) reception on the morphogenesis process applied by Tissue Engineering is a determinant of their structure which define their mechanical and biological properties. Tissue structures are built by the adaptive mechanism of interacting cells that respond to mechanical and biological signals. *In vitro* experiments that try to recreate biological structures have found several problems, particularly, fibrosis by excessive collagen synthesis. That is an important issue in tissues like heart valves, vessels, or ligaments. In previous works, an agent-based model was made to determine the emergent structure created by the interaction between mesenchymal stem cells stimulated with FGF-2, TGF-Beta1 Growth Factors, and mechanical signals to recreate a heart valve structure, but each line of code was programmed manually. In this work, the Model-Driven Approach (MDA) was applied with the purpose to create automatically executable models that permit simulating the tissue morphogenesis process. A Domain Specific Language (DSL) that modeled cells' behavior with their environment and considers gene activation has been defined. This DSL was represented in Eclipse Obeo using an Ecore metamodel. We focused on the use of Acceleo to apply Model To Text transformations for generating code of a multiagent system that runs in NetLogo.

Results: As a study case the morphogenesis heart valve models were made. A set of transformation rules from Model To Text were determined. Also, the recreated morphogenesis process corresponded with *in vitro* experiments done by other authors. Simulation models predict that TGF-Beta1 is the most important GF implicated in valve fibrosis doing a positive feedback loop.

Conclusions: MDA approach is a useful methodology to do Agent-Based models from DSLs. The defined Tissue Engineering DSL can be applied to simulate organs or tissues obtained *in vitro*. It predicts the influence of Growth Factor concentrations on tissue emergent structure. This DSL may be extended, and the executable code might be done for any other Multiagent environment applying transformation rules. In the next works, will be generated code for a 3D multiagent environment that permits the simulation of more complex tissues and organs.

POSTER #25

An inverse docking approach with biological networks to understand the effect of C. Citratus compounds on Chagas disease

J. Leonardo Gómez Chávez[1], Germán A. Conti[1], E. Rafael Perez[1], Emilio L. Angelina[1] , Nelida M. Peruchena[1]

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

Background: The chronic stage of Chagas disease is characterized by intense cardiomyopathy caused by infection by the parasite *Trypanosoma cruzi*. *Cymbopogon citratus* extracts have been shown to be effective at this stage, relieving this pathology by reducing amastigote nets and inflammatory infiltrates in the cardiac tissue of mice. In this work, bioinformatic tools together with chemoinformatic and virtual screening tools were used to understand the mechanism of action at the molecular level of *Cymbopogon citratus* extract in chagasic cardiomyopathy.

Results: The functional enrichment study of the GSE41089 dataset consisting of gene expression data in the heart of mice infected with *T. cruzi* versus control, resulted in a set of overexpressed genes (logFold-Change>1 and adjusted p-value <0.05) related to GO terms associated with the innate immune response and biological pathways of cytokine release. The Protein-Protein Interaction network built from the overexpressed genes in the most affected biological pathways highlighted those proteins with a high value of interactions in the network. Sequences of the prioritized proteins were aligned against the PDB database with the help of the BLAST engine to retrieve the proteins 3D structures. Native ligands in those structures were cross docked against all the remaining structures to evaluate and improve docking ability to match the known protein-ligand partners. Since selecting targets based on the raw docking scores has been shown to negatively impact the accuracy of reverse docking methods the Z-score metric was used to normalize ligands scores among all protein receptors. The calibrated reverse docking protocol was applied to screen each *Cymbopogon citratus* component against the prioritized proteins. Molecular Dynamics of the top-ranked complexes confirmed the inverse docking outcomes, including Nerolidol bound to Ptgs2 with a ΔG bind similar to specific inhibitors such as diclofenac and ibuprofen.

Conclusions: The combination of network pharmacology strategies together with molecular docking tools results in a promising strategy in the study of the action of plant extracts against Chagas disease and will allow the search for more effective treatments for it.

POSTER #27

Quercetin as a multitarget inhibitor: poor selectivity or something else?

Germán Andrés Conti, José Leonardo Gómez Chávez, Ernesto Rafael Perez, Emilio Luis Angelina, Nelida Maria Peruchena

Universidad Nacional del Nordeste

Background: Quercetin (QUE) is a naturally occurring phytochemical with several proven biological effects, including anticancer, antidiabetic, antiinflammatory, antioxidant, anti-Aging, antimicrobial, anti-Alzheimer's, antiarthritic, cardiovascular, and wound-healing effects. QUE excerpts those benefits by directly targeting key enzymes involved in several biological pathways. QUE can inhibit kinases, proteases, glucosidases, endopeptidases, hydrolases, topoisomerases, among others. A question that arises is: What makes QUE so versatile as to be able to bind and block such a diversity of molecular targets?

Results: To answer the above question we performed a structural survey with cheminformatic tools of QUE-enzyme complexes deposited on the Protein Data Bank (PDB). Alignment of all the QUE-enzyme complexes by the QUE atoms allowed the characterization of the environment of each QUE atom within the protein context. The extracted atomic environments were fed into cheminformatic and machine learning algorithms to expose the common pharmacophoric features that allow binding to all the different protein targets.

Conclusions: Analysis of QUE atomic environments provided some clues about a common enzyme inhibition mechanism by this flavonoid. Atomic environments could be also used to predict novel protein targets of QUE.

POSTER #29

Comparison of neural network-based methods for similarity prediction in compounds with unknown structure

E. Borzone, L. Di Persia, M. Gerard

Research Institute for Signals, Systems and Computational Intelligence, sinc(i), FICH-UNL, CONICET. eborzone@sinc.unl.edu.ar

Background: Similarity between compounds is widely used in chemoinformatics. Usually It is calculated using structural information of the compounds, so it is only available for compounds with known structure. To address this constraint, we use the information of the metabolic pathways topology, in order to infer similarity between compounds with unknown structure. In this work we compare on the same dataset three neural network-based models we have proposed, to solve the problem of compound similarity prediction.

Results: The first model we propose is a Multilayer Perceptron (MLP) with one-hot compound encoded inputs. It was used to explore if it was possible to predict similarity with an MLP. Although it can provide a good prediction of similarity, the used encoding makes it impossible to apply the model for compounds with unknown structure. The second alternative is an MLP for prediction, using as inputs embeddings obtained from a model of random walks through the graph of compounds of a metabolic pathway, that preserve the proximity of the compounds in relation to the reactions in which they participate. This model provided errors lower than 10% in test, and it has capabilities to predict similarity to compounds with unknown structure. Finally, the third proposal uses a message passing Graph Neural Network to generate embeddings for the compounds, and an MLP for the similarity prediction. The errors in the test set were close to 2%, producing a high performance model. The model can also preserve the topological properties in the generated embeddings, i.e., similar compounds are mapped to close embeddings in the embeddings space.

Conclusions: Three incremental models for similarity prediction of compounds with unknown structure are compared. From these promising results, we conclude that it is possible to predict similarity between compounds with unknown structures with good performance. In the future, features such as physicochemical information of the compounds will be incorporated to improve the generalization, and the models will be evaluated for larger metabolic pathways.

POSTER #30

Identification of molecular determinants involved in misfolded protein recognition by UGGT.

Juan Rodrigo Ortigosa[1,2], Rafael Betanzos San Juan[1,2], Marcelo Adrian Marti[1,2] and Carlos Pablo Modenutti[1,2]

1 CONICET - UBA. **2** IQUIBICEN

In eukaryotic cells, there is an efficient machinery in the Endoplasmic Reticulum (ER) that ensures the correct protein-folding of glycoproteins before their exit. UDP-glucose:glycoprotein glucosyltransferase (UGGT) is a soluble enzyme of 170 kDa that recognizes molten globe glycoproteins that exhibits hydrophobic residues and reglucosylates their high mannose glycan (Man9) in order to prevent premature secretion. This glucosylation promotes calreticulin, calnexin, and lectins association which enhances retention to reach correct folding or degradation in the ER. UGGT's structure has not been crystallized in humans yet. Recent structural and functional work from our group and collaborators was focused on Chaetomium thermophilum UGGT (CtUGGT) humans homolog which could be crystallized. Protein's multi-domain architecture has been uncovered and provided preliminary evidence of its inter-domain conformational flexibility. The full-length CtUGGT crystal structures revealed four DsbA-like domains (TRXL1–4) arranged in a long arc, terminating in two β sandwiches (βS1 and βS2) tightly clasping the glucosyltransferase family 24 (GT24) domain. Along with those results, it was proved that UGGTs lacking the TRXL2 domain showed impaired activity suggesting that the stated domain might be crucial for unfolded glycoprotein regions recognition. We utilize neural network-based analysis (NNBA) to predict protein-protein interaction regions. We obtained a significant score prediction for residues 560-580 from TRXL2. Along with these results, we performed molecular dynamics (MD) simulations with cosolvent (water + phenol) with the purpose of identifying hotspots or Solvent Sites (SS), a method able to map hypothetical interactions with core residues of misfolded proteins. Lately, we retrieve 53 non-redundant sequences of UGGT's homologs in Fungi's kingdom to inspect the conservation of surrounding regions of best-ranked SS. Considering the research we conclude that the two best ranked SS were nearby conserved regions. These latter are also putative protein-protein interaction regions according to our statistical analysis based on NNBA. In conclusion, we uncover the structural motifs which are essential in the recognition of molten globe glycoproteins. It is of great importance to point out regions that confer recognition to UGGT with the aim of finding new targets for drugs to modulate UGGT's behavior.

POSTER #31

Integrating multi-omic data using knowledge graph databases

Sergio Gonzalez[1], Diego Zavallo[1], Maximo Rivarola[1,2], Sebastian Asurmendi[1,2], Norma Paniego[1,2]

1 Instituto de Agrobiotecnología y Biología Molecular (IABIMO), INTA-CONICET, Hurlingham, Argentina. **2** CONICET, Buenos Aires, Argentina.

Background: The generation, consumption and especially the analysis of highly interconnected data have become ubiquitous. In this situation, where relationships between data increase in both quantity and importance, graph models become an appealing and viable solution, as graphs are mathematical entities in which objects are connected. In recent years, the explosion of densely interconnected, heterogeneous, multi-omic data has promoted the research, development and adoption of integration technologies. Due to the interconnected and heterogeneous nature of its data, the field of multi-omics can be one of the beneficiaries of graph databases, which enable more natural representation models, better data integration workflows and exploratory analysis. In this work, we present the development and implementation of a set of programs to load genomic data into a graph database in an integrated manner.

Results: We define a data model that consists of deciding how to bring information into the graph in the form of vertices and edges. For example, in this model every genomic variation, allele and genotype are vertices and the presence of a particular allele in a genotype is an edge between them. Based on this modeling, we implement a set of Python scripts to load structured genomic information into a graph database based on the Neo4j platform. We developed a command line interface to take information in several structured text formats, such as, GFF3 (with and without hierarchical annotation structure), VCF, GAF, etc. We also incorporate different ontologies, based on GO and Reactome, and develop a general script that allows loading any ontology in OBO format. We use Arabidopsis thaliana genomic information as case of use, and load 27655 genes, 54013 transcripts (mRNA and ncRNA), 12883871 genomic variations (SNP, indels, etc) with associated genotyping information, 43564 GO terms and the complete Reactome DB.

Conclusions: We conclude that our graph database management system is suitable and efficient as a knowledge representation to support intensive and strongly connected data. Integrative analysis is aimed for both, basic research as well as applied agricultural problems (i.e. plant breeding, plant-stress responses). Future steps include using the database to address large scale complex queries based on expected relationships, and using graph algorithms such as path finding, clustering, vertex centrality and edge prediction, to find non-expected relationships in the dataset or to explore graph structure.

POSTER #32

Bioinformatic pipeline for protein binding site characterization

Marcelo Daniel Gamarra[1], Juan Blanco[1] and Carlos Modenutti[2]

1 FCEN-UBA. **2** University of Buenos Aires

In recent years, structural bioinformatics has played an important role in predicting structure and estimating protein function. Molecular Docking algorithms allow to predict protein–ligand structures, and there are several packages available in the community, also Molecular Dynamics estimate how they behave in aqueous solution through time. These tools are useful for identifying the position of molecules at their target binding sites and in turn revealing the main amino acids of the interaction. Phage J-1 belongs to the *Siphoviridae* family and can infect many strains of *Lactobacillus casei/paracasei* used in elaboration of fermented products, either retarding production, affecting quality of the product or totally interrupting the process. Carbohydrate Binding Module 2 (CBM2) of "Dit" protein present in Phage J-1 base plate, is directly involved in the recognition of bacterial cell wall carbohydrates. In this context, we present here a pipeline that allows identifying the key amino acids for protein-ligand interaction, applying it to the recognition of molecular determinants of phage J1 - *Lactobacillus casei* interaction. For this work, free-use bioinformatic tools were used. BLAST protein, DALI server and Fpocket were used for the identification of domains and binding sites. Autodock4.2, AmberTools21, Amber were used for docking and molecular dynamics calculations. Computational tools including DM and Docking assays allowed to identify three amino acids (H410, W440 and D498) that show a polar interaction more frequently with all variants of the ligand. Rhamnose saccharide present in all ligands showed the highest frequency of receptor interaction according to experimental data. In this sense, the results obtained show that the combination of web available tools for use with computational simulation methods of molecules (docking, DM and solvent sites interactions) are capable of detecting the key amino acids in the interaction of the protein with its target carbohydrate. The proposed method was evaluated *in vitro* demonstrating its validity. The results obtained show that the proposed pipeline is useful not only for knowing how a ligand binds to its target protein, but also how the recognition process is. All this contributes to the molecular knowledge of the phage-bacteria interaction, directing the development of strategies that prevent phage infection.

POSTER #34

Performance evaluation of Gyra for the taxonomic classification of the clade Bacillus subtilis using a machine learning approach

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

1 IPROBYQ-CONICET, Rosario, Argentina; 2 FCByF-UNR, Rosario, Argentina;CEFOBI-CONICET, Rosario, Argentina; 4 LIFiBVe, ICiAgro Litoral, UNL, CONICET,FCA, Santa Fe, Argentina, petittitomas@gmail.com

Background: Due to the diversity of the B. subtilis clade, multiple cases of inconsistencies between taxonomic classification and genomic or phenotypic characteristics or incorrect classifications have been reported. However, the proper assignment is critical since these are used to estimate safety and performance of bacteria, impacting its use in industry and agriculture.

Results: In order to correct the taxonomic classification of genomes of the clade B. subtilis, 2625 sequences belonging to the clade were downloaded. Then, 133 low quality sequences were eliminated. The taxonomic identity was validated or reassigned using Average Nucleotide Identity and multi-locus sequence analysis. Thus, 29.5% of the sequences were reassigned. In turn, 148 strains were classified into 12 new genomospecies. The K-mers of the gyrA gene sequences were obtained. Classification models based on RF and SVM with linear and radial kernels were generated. The predictive capacity of these models was evaluated using K-Fold repeated CV. For the RF algorithm, a tuning grid of hyperparameters were evaluated, obtaining values of Kappa between 0.9968 and 0.9984. Using the testing set, Kappa values between 0.993 and 1.0 and values of AUC = 1 were obtained, verifying that the models do not perform overfitting. The linear kernel and the radial kernel were evaluated with different hyperparameters. The linear kernel for the best model (c = 2 0.5) showed an error of 0.1% for the training set; for the testing set Kappa and AUC were 0.9969 and 0.9995, respectively, indicating that the model does not perform overfitting. The best model using the radial kernel showed an error of 17.7 % for the training set. The testing set showed a Kappa of 0.472 and an AUC of 0.82.

Conclusions: These results indicate that data is linearly separable, and that RF is the best algorithm for rapid and massive classification. From these results, the development of a tool that allows the classification of metagenomes within the B. subtilis clade, with similar levels of resolution and less time required than whole genome analysis, is proposed.

POSTER #37

Evaluation of four genome assembly tools for third-generation PacBio long-read sequence data analysis to obtain a high-quality de novo genome from Verticillium dahliae

Pablo Aguilera, Ben Guerrero, Sergio Gonzalez, Maria Carolina Martinez and Norma Paniego

INTA

Background: Leaf mottle and wilt caused by the soil-borne ascomycete fungus Verticillium dahliae Kleb., is one of the most important diseases of sunflower. Two local phytopathological races, VArg1 and VArg2, have been described in Argentina. The reference genome of *V. dahliae* is 35Mb in size and consists on 8 chromosomes. In this study, we obtained longread hifi PacBio sequencing for the genomic analysis of two Argentine *V. dahliae*-pathogenic sunflower isolates to obtain a high-quality de novo genome. For this purpose, we evaluated 4 HiFi assemblers for Pacific Biosciences (PacBio) reads.

Results: Sequencing of two *V. dahliae* genomes using two PacBio Sequel II SMRT cells yielded 4.5 and 2.8 Gb of HiFi data, corresponding to ~128x and ~81x genome coverage, respectively. These data correspond to 767k and 353k reads with a mean length of 5922bp and 8058bp, respectively. PacBio datasets of *V. dahliae* were analyzed using HiCanu, Hifam, Flye, and IPA assemblers. The resulting assemblies contained between 20 and 293 contigs with an N50 between 646kb and 4147kb and the assembly size range 34-38 Mb, as expected. To check the quality of each assembly, we compute several metrics. First, we analyze the Kmer distribution to check the coherence of the assembly based on the content of the reads with the software KAT. The QV score of the Inspector software ranged from 45 to 57. The analysis of the Merfin software revealed QV * scores ranging from 19 to 21.63 and completeness scores ranging from 0.97 to 0.98. Considering all metrics evaluated, the Flye assembler performed better on Varg2 sequences, while the IPA assembler performed better on Varg1 sequences. A final analysis using Tapestry software allowed us to assign telomeres to these assemblies.

Conclusions: Comparing the performance of different assemblers based on different metrics is essential to generate a high quality assembled genome in non-model organisms. The genome assemblies of the two *V. dahliae* isolates obtained here are complete chromosomal scaffolds. This quality promises a comprehensive study of the genomic regions that confer the ability to cause vascular wilt disease in sunflower, as well as further studies on the interaction of this fungus with its host.

POSTER #40

Diagnosis of human aneuploidies, discrimination of SARS-COV-2 variants and determination of animal species in Halal foods through high-resolution melting curves previously evaluated by different in silico tools.

Pablo Vélez[1], Guillermo Giaj-Merlera[1], Jésica Gallardo[2], Melanie Neira[1], Félix Condat[3], Lorenzo Rosales[4], Andrea Belaus[1]

1 Molecular Biology Unit. Ceprocor. Ministry of Science and Technology of the Province of Córdoba. **2** Food Unit. Ceprocor. Ministry of Science and Technology of the Province of Córdoba. **3** Department of Organic Chemistry, Faculty of Chemical Sciences, National University of Córdoba. **4** Department of Microbiology and Immunology. Faculty of Exact, Physical-Chemical and Natural Sciences. National University of Rio Cuarto.

Background: The real-time PCR reaction followed by high resolution melting curves (HRM) provides a fast and low-cost alternative for the discrimination of amplicons under study. From the DNA sequence, the theoretical melting temperature (Tm) of the double strand can be known using thermodynamic tables that are incorporated into different bioinformatics tools. The BioPython "MeltingTemp" module and the UMelt web tool were used to determine the theoretical Tm of candidate regions. As an objective, it was proposed to find and validate candidate regions whose in silico simulations present differences in theoretical Tm (ΔTm) useful for discrimination, by means of the real-time PCR test with HRM. These analyzes were performed in three different areas of interest.

Results: In Human Health, 14 regions of genome segment duplications met the amplification requirements and diagnostic potential for common human aneuploidies (21, XY, 18, and 13). A 106 bp segment of the duplication UID-15370 (for Trisomy 21) obtained in vitro a ΔTm of 2.6°C, observing specificity and consistency between samples from affected and unaffected individuals. The remaining 13 regions continue to be studied. Regarding the monitoring of variants of SARS-COV-2, the viral genomes were downloaded from public databases, then aligned and made consensuses, later the differences were searched for amino acid changes reported in the variants defining optimal regions according to their theoretical Tm. On 24 positive clinical samples (confirmed by sequencing), two different regions of the S gene of the virus made it possible to separate the variants associated with Gamma and Lambda from the rest of the variants. For discrimination in Halal-certified foods, the 12S and 16S ribosomal RNA genes were chosen by manual inspection over alignments of 17 reference sequences from animals of regional interest. Both genes combined allowed to separate 13 species in silico.

Conclusions: Simulations of curves and dissociation temperatures allow to determine candidate regions of the DNA with potential discrimination by real-time PCR followed by HRM based on theoretical ΔTm. The in vitro results were consistent with the in-silico analyses.

POSTER #41

multiPAML2.5: a Python script to run multiple molecular evolution PAML analysis

Francisco Pisciottano, María Clara Campos and Patricia Saragüeta

Instituto de Biología y Medicina Experimental - CABA, Argentina

Background: D_N/D_s positive selection analysis is a widespread approach to search for signatures of molecular evolution and to identify genes involved in evolutionary events. Among the available software to perform such analyses the package of programs PAML (Phylogenetic Analysis by Maximum Likelihood) is the most used to construct Likelihood Ratio Tests of evolutionary hypotheses.

Results: Here we present multiPAML2.5, a simple homemade Python script that was designed to run multiple PAML branch-site positive selection analyses. multiPAML2.5 was designed to help the user deal with the tedious PAML process required to run positive selection tests on multiple genes. The script considers multiple evolutionary hypotheses with minimal manual user intervention as it handles directory organization, files distribution, program execution, results gathering of Likelihood Ratio Test. multiPAML2.5 was implemented and tested in the evolutionary analysis of gamete interaction proteins along the Carnivora order. Several oocyte (ZP1, ZP2, ZP3, ZP4, IZUMO1R) and sperm (ZP3R, ZAN, SLLP1, PKDREJ, IZUMO1) proteins were analyzed from highly curated MSAs. multiPaml2.5 allowed us to test multiple evolutionary hypotheses of both episodic and divergent positive selection in the basal branch and the sub-trees of Carnivora, Caniformes, Feliformes and Pantherines for the listed genes in just one scrip run, rendering an organized directories hierarchy and a single simple but efficient output results file.

Conclusions: multiPAML2.5 proved to be a handy and accessible way of running numerous PAML analyses from multiple MSAs corresponding to different genes and to test multiple evolutionary hypotheses for each of them in just one script run sparing the user the trivial chores required to set a PAML analysis and gather results which are error-prone and time-consuming when performed manually. In the example of the gamete interaction proteins of the Carnivora order showed here, the results will be useful to understand the molecular basis of the lack of precygotic reproductive isolation mechanisms in felids and compare them to canids, its sister group.

POSTER #42

Genomic and proteomic characterization of HPV-16 and HPV-31: A comprehensive bioinformatic analysis of these two genotypes that prevail in fertile women from the city of Posadas, Misiones.

Florencia Benitez[1,2], Carolina S. Cerrudo[3], Silvina E. Hanke[1], Mariano N. Belaich[3], Verónica Martínez Marignac[2] and Graciela B. Jordá[1]

1 Laboratorio de Microbiología Especializada, Módulo de Farmacia y Bioquímica, Facultad de Ciencias Exactas, Químicas y Naturales (FCEQyN- UNaM). Posadas – Misiones. **2** Laboratorio Interdisciplinario de Biología y Genética Molecular (IBIOGEM) Centro de Investigaciones Científicas y Transferencia de Tecnología a la Producción (CICYTTP). Diamante – Entre Ríos. **3** Laboratorio de Ingeniería Genética y Biología Celular y Molecular – Área Virosis de Insectos (LIGBCM-AVI), Instituto de Microbiología Básica y Aplicada (IMBA), Departamento de Ciencia y Tecnología, Universidad Nacional de Quilmes. Bernal - Provincia de Buenos Aires.

Background: The human papilloma virus (HPV) and *Chlamydia trachomatis* are the most frequent causes of sexually transmitted infections in adolescents and young people with risk factors, particularly in women under 25 years of age. Both infections can cause sexual and reproductive health consequences, and some HPVs can even cause invasive cancer. The *Papillomaviridae* family, a diverse group of double-stranded DNA viruses with a protein capsid, consists of more than 450 genotypes, and about 40 of them can infect the genital tract. Epidemiological studies of global prevalence have shown that these HPV types can be subdivided into two groups based on their oncogenic potential: low-risk and high-risk. The two genotypes that, according to our studies, prevail in fertile women in the city of Posadas (Misiones, Argentina) are HPV-16 and HPV-31, both of which belong to the high-risk group and, in the form of persistent infection, can lead to cancer.

Results: In this sense, we performed a bioinformatic analysis that includes the phylogeny of the complete genomes of HPV-16 and HPV-31, and an exhaustive characterization of the proteins that encode. The results could allow us to improve our knowledge about these viral genotypes and their relationship. Phylogenetic and molecular evolutionary analyses were conducted by using IQ-TREE 2.1.2, the identification of protein motifs and domains were performed using InterProScan, secondary structure were predicted with JPRED 4 server, and homology modeling protein structures were constructed using I-TASSER server.

Conclusions: In this work we present an update of the current knowledge about HPV-16 and HPV-31 genotypes and by means of a comprehensive bioinformatics analysis we characterized their proteomes and genomes to expand the knowledge of their evolution, secondary and tertiary structure, and function of their proteins, contributing to future research in the area of HPV.

POSTER #43

T cell immunogenicity prediction of peptides presented in MHC class I molecules

Ibel Carri[1], Heli Magalí García Alvarez[1], Morten Nielsen[2]

1 National University of San Martín, **2** Denmark Technical University of Denmark

Background: Epitopes are defined as the region of an antigenic protein recognized by the immune system. Epitope identification is relevant to the study of immune responses and is crucial for vaccine development. In the immune response mediated by CD8+ T-cells, one of the most critical steps to define T-cell recognition or immunogenicity is peptide-MHC binding. Approximately, 1 in 200 peptides are recognized by the MHC complex and presented on the cellular surface. For this reason, multiple predictive tools have been developed to predict MHC binding with relative success. However, to elicit an immune response, the peptide not only needs to be presented on the cell surface but also be identified by the corresponding T-cell. For this reason, relying on peptide binding is not enough to efficiently predict immune response, especially when exploring large and complex proteomes such as parasites or cancer. The scientific community has identified that the aminoacidic composition of peptides influences immunogenicity, but predictive tools based on this information have limited performance and epitope prediction remain an unsolved challenge. This is mainly due to the combination of data scarcity and the complexity of the mechanism of T-cell recognition. In this work, we propose a novel machine learning-based tool that integrates the peptide sequences with several calculated features from the peptide and its source protein to improve immunogenicity prediction.

Results: We obtained experimentally validated epitopes, non-epitopes, and their source proteins from the Immune Epitope Database. Only peptides with good binding to MHC were included to prevent the model to learn this known related feature alone. We included epitopes without the cognate MHC allele information and reconstructed the MHC specificity to enlarge the dataset. Proteic and immunological features were calculated based on peptide and protein sequences of this dataset. Proteic features, especially the ones related to protein structure, were demonstrated to correlate with immunogenicity. Different combinations of said features will be combined in a convolutional neural network.

Conclusions: The analysis presented indicates that peptide and protein sequences contain complementary information related to determinants of immunogenicity, which combined may improve immunogenicity prediction and reveal rules defining T cell recognition and activation.

POSTER #45

Comparative plastomic among populations of the species that originated the cultivated peanut, Arachis hypogaea

Samoluk SS[1,2] , Moreno EMS[1,2], Abernathy B[3], Bertioli DJ[3], Seijo JG[1,2]

1 Instituto de Botánica del Nordeste, Corrientes, Corrientes, CP 3400, Argentina. **2** Facultad de Ciencias Exactas y Naturales y Agrimensura (Universidad Nacional del Nordeste), Corrientes, Corrientes, CP 3400, Argentina. **3** Center for Applied Genetic Technologies, University of Georgia, Athens, GA, USA.

Background: The elucidation of the genetic and geographic origins of cultivated species is important for the conservation and utilization of new genetic variation. The cultivated peanut (Arachis hypogaea) is an allotetraploid species with "AABB" genome constitution. *Arachis ipaensis* is the species that donated B subgenome of peanut as a paternal parent. *Arachis duranensis* is the maternal donor (A genome), however, there are still some controversial issues regarding: 1) the population/s that participated as donors of the A subgenome; and 2) the single and recent (<10,000 years ago) origin of peanut, involving a wild allotetraploid ancestor (*A. monticola*), versus a multiple and much older origin (~450,000 years ago) for the cultigen. Here we analyzed the plastome variability among accessions of *A. duranensis*, *A. monticola*, *A. hypogaea* and outgroup species aiming to shed light on these controversies.

Results: The analysis of a 9572 bp chloroplast SNP alignment from 95 wild and cultivated accessions identified 46 haplotypes, 11 of them belonging to outgroup species. All other haplotypes belong to *A. duranensis*, *A. monticola* and peanut arranged (except four) into three haplogroups in the network (A, B and C). Haplogroups B and C were exclusively composed of *A. duranensis* accessions. The haplogroup A, the largest one, included all accessions of *A. monticola* and *A. hypogaea*, and some accessions of *A. duranensis*. This haplogroup showed a central haplotype shared by accessions of these three species, and several derived species-specific tip haplotypes, all of them separated by one to four mutational steps. All *Arachis hypogaea* accessions showed a narrow plastome diversity without a clear segregation between the botanical varieties and subspecies.

Conclusions: Despite the results are not enough to identify the population from which peanut may have originated, they revealed a rapid, unique and recent diversification of *A. hypogaea* from an ancestor that belong to haplogroup A.

POSTER #46

Automatic prediction of cervical cancer from associated risk factors

Anibal Chelaliche[1], Cintia Kaufman[2], Elizabeth Tapia[3], Flavio E. Spetale[3]

1 Laboratorio de Biotecnología Molecular, FCEQyN - UNaM. **2** Instituto de Inmunología Clínica y Experimental de Rosario. **3** CIFASIS.

Background: Cervical cancer is the fourth most common neoplasm in women worldwide. For some time, a design has been pursued for a classification system that can be used in the diagnosis of this pathology using biochemical parameters, demographic data, environmental, genetic factors and images of the cervix.

Results: In this work we present an Artificial Neural Network (ANN) to predict the risk factors of CC. In addition, the Support Vector Machine Recursive Feature Elimination (SVM-RFE) method was used to find the most important attributes for cervical cancer prediction. The dataset employed here contains missing values and is highly imbalanced. Therefore, the Tomek Links technique was employed to remove this data from the majority class (negative examples). The feature-selection method showed that the frequency of tobacco consumption, HVP and HIV infections, as well as the number of diagnosed sexually transmitted infections (STDs), could be useful for predicting cervical cancer. The ANN with the selected features from SVM-RFE and Tomek Links showed the best results with an accuracy of 78 %.

Conclusions: The associations among certain oncogenic strains of HPV, others STDs and tobacco consumption, and the disease are well established. The development of this model applied to the clinic can become a useful tool for the first stages of patient screening, which helps health professionals to choose subsequent methodologies for disease control and treatment.

POSTER #47

Validation and finding of distant orthologous genes of the DAF-12 gene in Meloidogyne Incognita using bioinformatics tools.

Rafael Betanzos San Juan[1,2], Claudio David Schuster[1,2] and Carlos Modenutti[1,2]

1 Departamento de Química Biológica, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Ciudad Universitaria, Pab. II (CE1428EHA), Buenos Aires, Argentina. **2** Instituto de Química Biológica de la Facultad de Ciencias Exactas y Naturales (IQUIBICEN) CONICET. Ciudad Universitaria, Pab. II (CE1428EHA), Buenos Aires, Argentina.

Root-knot nematode (RKN) Meloidogyne spp. is one of the most damaging parasites due to its wide range of hosts. Here, we report a DAF-12 homologue from Meloidogyne incognita (DAF-12Minc) that had not been described before and that is a promising target to control infestations. Using a combination of Hidden Markov Models (HMM) based sequence search and phylogenetic analysis we identify three orthologous of *C. elegans* receptor DAF-12 in M. incognita. These candidates displayed high identity percentages between themselves, indicating that it could come from a triplication event. Despite the global sequence identity between previous DAF-12 orthologous and DAF-12Minc were acceptable, the key amino acids in the hormone receptor binding site were missing. In that sense, we evaluated if our DAF-12Minc candidates were able to bind Dafachronic acids (DA's) by modeling the structure of DAF-12Minc ligand-binding domain and performed Molecular Dynamics to refine the models, pocket detection to identify the ligand binding site of the protein and molecular docking to evaluate previously reported ligands. These analyses suggested that the features of the active site are conserved. Moreover, the luciferase reporter gene led us to elucidate the antagonistic effects of the DA's; and further *in vitro* and *in vivo* analysis suggest that DA's had an effect in both life cycle and mortality of *M. incognita*, produces an overall improvement in plant health condition due to the effect of the compound on *M. incognita* life cycle, and that DA's could have an effect not only in the infection by *M. incognita*, but also a long-term effect in the inhibition of offspring hatching.

POSTER #49

Developing an Automated Protocol for the Wristband Extraction Process Using Opentrons

Tei Kim[1], Brooklynn McNeil[2], Kathryn Dunn[2], Douglas I. Walker[2,3]

1 Stanford Online High School, 415 Broadway, Redwood, City, CA. **2** Department of Environmental Medicine and Public Health, Icahn School of Medicine at Mount Sinai, New York, NY. **3** Gangarosa Department of Environmental Health, Rollins School of Public Health, Emory University, Atlanta, GA

To better characterize the relationship between complex chemical exposures and disease, our laboratory uses an approach that combines low-cost, polydimethylsiloxane (silicone) wristband samplers that absorb many of the chemicals we are exposed to with untargeted high-resolution mass spectrometry (HRMS) to characterize 1000's of chemicals at a time. In studies with human populations, these wristbands can provide an important measure of our environment: however, there is a need to use this approach in large cohorts to study exposures associated with disease. To facilitate the use of silicone samplers in large scale population studies, the goal of this research project was to establish automated sample preparation methods that improve throughput, robustness, and scalability of analytical methods for silicone wristbands. Using the Opentron OT2 automated liquid platform, which provides a low-cost and open source framework for automated pipetting, we created two separate workflows that translate the manual wristband preparation method to a fully automated protocol that requires minor intervention by the operator. These protocols include a sequence generation step, which defines the location of all plates and labware according to user-specified settings, and a transfer protocol that includes all necessary instrument parameters and instructions for automated solvent extraction of wristband samplers. These protocols were written in Python and uploaded to Github (https://github.com/teikimm307/wristbandautomated) for use by others in the research community. Results from this project show it is possible to establish automated and open source methods for preparation of silicone wristband samplers to support profiling of many environmental exposures. Ongoing studies include deployment in longitudinal cohort studies to investigate the relationship between personal chemical exposure and disease.

POSTER #50

Identifying conservative sites of Sortase A protein in different strains of Enterococcus faecalis

Mauricio González[1], Silverio Andrés Quintana[1]

1 Universidad Nacional de Asunción, Facultad de Ciencias Exactas y Naturales, Departamento de Biotecnología

Sortase A is a transpeptidase from Gram-positive bacteria which catalyzes cell-wall sorting of proteins to the peptidoglycan cell wall by cleaving a LPXTG motif between the threonine and glycine, and later binding the carboxyl group to a pentaglycine in the cell wall. The activity of the sortase family protein was discovered in the early 1990s by Schneewind and the first sortase isolated was the Sortase A from Staphylococcus aureus. Computational studies of the sortase-peptide complexes provided the initial interaction between the sorting signal to the active site and formation of the first tetrahedral and thioacyl intermediates. The protein changes formation so that the active site cysteine from the TLITC, part of the catalytic and ligand-binding sites, cleaves the guanine in the LPXTG motif. Currently, there are 4 classes of Sortase, divided by the recognition sequence, and most Gram-positive bacteria have at least Sortase A, which is used for sorting adhesion and immune evasion factors. The microorganism used in this study is *Enterococcus faecalis*, a known commensal bacteria inhabiting the gastrointestinal tracts of humans which can be used as probiotic, but can cause urinary tract infections and endocarditis. The main purpose of this study is to identify conservative sites in Sortase A through alignment of sequences taken from NCBI (RefSeq WP_161323801.1 from 132 to 256AA) using Blastp and them ClustalW. Most strains like Merz96, TX2141 and ATCC 29200 presents 100% identity while others, like TX1467 present at least 61% identity as a result of alignment with Blastp and it was also observed a 100% conserved site of TLITC. Subsequent identification of these sites in an structural analysis of proteins from the Refseq and different strains of *E. faecalis* found in different environments, can be later used as targets for pharmaceutical drugs for pathogenic strains and biotechnological applications such as cost-effective ligation of proteins or substrate conjugation. In conclusion, it can be mentioned that the sortase A protein has conserved sites across the majority of different strains *E. faecalis* used in the alignments, maintaining a high degree of identity in the amino acids TLITC.

POSTER #51

Using genomic sequence variability to reveal PRMT5-mediated splicing regulation

Maximiliano Beckel[1], Andres Rabinovich[1], Marcelo Yanovsky[1], Ariel Chernomoretz[2]

1 Fundación Instituto Leloir, **2** University of Buenos Aires

Background: Protein arginine methyltransferase 5 (PRMT5) regulates several molecular processes, including transcription, cell signaling, DNA damage response, chromatin modification and splicing. PRMT5 methylates some SR proteins and core components of spliceosome, such as, Sm and LSm proteins. PRMT5 inhibition produces remarkable changes in splicing patterns. It has been proposed that the action of PRMT5 could favor the recognition of weak donor sites.

Results: In order to determine to what extent the effect PRMT5 has on splicing patterns is influenced by cis signals, sequencing experiments were performed (RNA-Seq) in PRMT5 wild type and mutant Arabidopsis thaliana plants belonging to two different ecotypes: Columbia (Col-0) and Landsberg erecta (Ler). In this way, we sought to analyze the effect that the lack of PRMT5 activity has on the genetic variability that these accessions have. Aiming to distinguish the changes related to variations in the cis sequences, F1 Col-0 X Ler and Ler X Col-0 hybrids were also analyzed. We found that in both accessions there is an increase in the expression of a large number of factors related to the regulation of splicing and the spliceosome when PRMT5 is absent. On the other hand, a greater number of splicing events are altered due to the sequence differences present between the accessions. When comparing the percentage of intron retention that occur in events that present sequence differences in the 5'ss, we demonstrated the importance of the strength of these sites for determining the effect of PRMT5 on splicing. The analysis of the effect of the absence of PRMT5 in the case of hybrid plants, allowed to determine that at least 42% of the differential splicing events found are due to the differences in sequences that occur between one accession and another.

Conclusions: Through these analyses, it was concluded that an important part of the splicing patterns affected by the PRMT5 mutation depend on cis signals, having a particular relevance the strength of the donor site. These results allow us to establish that PRMT5 may be an important factor in maintaining the fidelity of the splicing process in the face of changes, such as mutations in the genomic sequences associated with splicing.

POSTER #52

An in silico approach to characterize the resistance against heavy metals and UV-B radiation of poly-extremophilic Nesterenkonia strains

Natalia N. Alvarado[1], Nicolás Palopoli[2], Virginia H. Albarracín[1]

1 Centro Integral de Microscopía Electrónica (CIME), CCT-CONICET, UNT, Tucumán, Argentina. **2** Departamento de Ciencia y Tecnología, Universidad Nacional de Quilmes - CONICET, Bernal, Buenos Aires, Argentina.

Background: *Nesterenkonia* is a genus of halophilic actinobacteria that thrive in extreme environments such as the High Altitude Andean Lakes (HAAL) in the South American Central Andes. These microbial communities present unique mechanisms to adapt to adverse conditions such as high levels of ultraviolet radiation, hypersalinity, desiccation, high concentration of heavy metals, and temperature fluctuations. We previously demonstrated that the UV-B and heavy metals resistance profiles of *Nesterenkonia* Act20 (Act20), isolated from HAALS, are superior than those of *Nesterenkonia halotolerans* YIM70084 (NH), a reference strain from high salinity soils in China. Here we compare computational genome-wide annotations of Act20 and NH, previously obtained in our group, with other reference organisms to pinpoint candidate genes that could explain the resistance (or lack thereof) to stress in *Nesterenkonia*. Sequence-based analysis of amino acid residue conservation provides further details about the best candidate proteins.

Results: We found that high UV-B resistance could be related with differences in CPD photolyases, a class of light-dependent enzymes that repair DNA damage caused by formation of cyclobutane pyrimidine dimers induced by ultraviolet light. We observed that the predicted Act20 CPD photolyase gene is 48.88% identical to the well-described CPD photolyase PhrA of *Agrobacterium tumefaciens*. Moreover, almost all of the amino acid residues that constitute the binding site for 5,10-MTHF as an antenna chromophore, as well as those that bind DNA and FAD, are conserved at the protein level. In contrast, the putative NH CPD photolyase has a lower identity of 22.69% with PhrA, and barely any conservation of functional residues.

Conclusions: Although Act20 and NH might be very similar (i.e., their 16 S-rDNA sequences share 97.4% identity), we found significant differences in coding genes that may account for the observed stress responses. We propose that a CDP photolyase has an essential role for the high UV-B resistance in *Nesterenkonia* Act20 that is absent in NH. Experimental validation will allow us to test this hypothesis. We also observed suggesting differences between Act20 and NH in genes that mediate copper sequestration, although similar computational analysis are needed to characterize the heavy metal resistance profiles.

POSTER #53

Identification of SIAH E3 ligase partners

Vera Ujvári, Tamás Szaniszló, Gábor Erdős, Mátyás Pajkos, Zsuzsanna Dosztányi

1 Department of Biochemistry, Eötvös Loránd University, Budapest, Hungary

E3 ligases are key enzymes that play a critical role in proteostasis by ubiquitinating and targeting their substrates for proteasomal degradation. These enzymes mostly recognize their interacting partners through short linear motifs (SLiMs). Due to the rapid degradation process, the transient, low-affinity nature of the interactions and due the difficulty of identifying proteins from poorly soluble cellular compartments, the identification of substrates has been challenging. BioID is a proximity-dependent biotin labeling technique, which allows the identification of low affinity, transient ligase-substrate interactions. Using BioID with 26S proteasome inhibitor (MG132) and analyzing the results of mass spectrometry, we are planning to identify new substrates for Seven in absentia homolog (SIAH) E3 ligases. SIAH proteins are E3 ligases with well characterized biological functions including tumor suppression, specifying cell fate, apoptotic cell death, transcription regulation, spermatogenesis and axon guidance. However, currently only a few of their substrates have been identified. Combining bioinformatic predictions with the results of BioID, our goal is to elucidate the unknown part of the interacting network of these enzymes. Identified hits are validated by biochemical methods, including GST pull down, co-immunoprecipitation and confocal microscopy. For the characterization of the interactions, measuring binding properties and validating SLiMs, we are going to use fluorescence polarization and isothermal titration calorimetry. With the results of these experiments, we are going to broaden our knowledge about SIAH E3 ligases and create a well-established methodology for E3 ligase interactome mapping. Through the understanding of these protein-protein interactions and their possible malfunctions in detail, these results can provide a basis for therapeutic research.

POSTER #55

Disclosing hidden metabolic traits in plant-fungal interactions: Identification of hub metabolites by correlation network analysis

Amira Susana Nieva[1,2,3,4], Fernando Matías Romero[3], Alexander Erban[1], Pedro Carrasco[5], Oscar Ruiz[3], Joachim Kopka[1]

1 Max Planck Institute of Molecular Plant Physiology (MPI-MP), Am Mühlenberg 1, 14476 Potsdam, Germany. **2** Postdoctoral Fellow 2019-2021 - Deutscher Akademischer Austauschdienst (DAAD), Kennedyallee 50, 53175 Bonn, Germany. **3** Instituto Tecnológico de Chascomús (INTECH), Universidad Nacional de San Martin (UNSAM), Av. Intendente Marino Km 8.2, Chascomús 7130, Argentina. **4** Current address: Cátedra de Fisiología Vegetal, Facultad de Ciencias Agrarias, Universidad Nacional de Catamarca (UNCA), Av. Maestro Quiroga 80, Catamarca 4700, Argentina. **5** Institut de Biotecnològia i Biomedicina (BIOTECMED), Universitat de València, Av. Doctor Moliner 50, 46100 Burjassot, Spain.

The forage legume *Lotus tenuis* represents an important source in constrained environments due to its capacity to tolerate abiotic stress conditions. *L. tenuis* has been correlated with Fusarium species which are key pathogens of important crops. However, some *Fusarium* species including the *F. solani Species Complex*, establish endophytic interactions with legumes. In this trend, *Fusarium solani* manages to infect the tissues of *L. tenuis* and *L. japonicus*. Preliminary metabolomics analysis revealed the involvement of phosphorylated compounds as metabolic traits of these interactions. However, the statistical approaches based either on one-way or multivariate analysis lack the systemic view denoted by the interaction between variables. We conducted metabolomics analysis based on gas chromatography-mass spectrometry (GC-MS/EI-TOF) on plant tissues under optimal conditions (Control), severe phosphate starvation (-P), F.solani presence (FUS+) and stress combination (FUS+P-). We profiled and analyzed the primary metabolism by conventional statistical approaches and took a step beyond by including a comparative correlation network analysis as a strategy to detect changes in interactions between central metabolites in response to our experimentally defined conditions. Considering the correlation network approach, the detection of patterns among metabolite interactions may change in response to the adaptation or acclimation of plant metabolism to environmental conditions. Such changes in interactions are typically not revealed by analyses of individual metabolites and they can exist even between metabolites that do not share the same metabolic pathway. By analyzing the network topology, we follow a pipeline to determine “hub” metabolites for each biological system. Therefore, these “hub” metabolites are consequently assumed to play crucial biological roles. Our results demonstrated that most hubs were present in the FUS+P- *L. japonicus* shoot network. The subsequent classification of compounds according to degree classes and betweenness centrality parameters determined a suit of hubs represented by biotic and abiotic stress-related metabolites. Among our results, sugars and sugar-related compounds exhibited relevant alterations in the metabolic response to the FUS+P- combination. Taken together, the determination of hubs by metabolite network analysis demonstrates the potential of this approach to find new perspectives in plant-fungal interactions research.

POSTER #56

Identification and functional analysis of differentially expressed genes in roots of drought-stressed Yerba Mate plants

Edgardo Hernán Avico[1], Raúl Maximiliano Acevedo[1], María José Duarte[1], Oscar Adolfo Ruiz[2], Pedro Alfonso Sansberro[1]

1 Instituto de Botánica del Nordeste (IBONE), UNNE-CONICET, Corrientes, Argentina. **2** Instituto de Investigaciones Biotecnológicas - Instituto Tecnológico de Chascomús (IIB-INTECH), UNSAM-CONICET, Chascomús, Argentina.

The appearance of water stress episodes triggers leaf abscission and decrease *Ilex paraguariensis* yield. To explore the mechanisms that allow it to overcome dehydration, we investigated how gene expression variation between water-stressed and non-stressed plants and in what way the modulation of gene expression was linked to physiological status and metabolite composition. A total of 5160 differentially expressed transcripts was obtained through RNA-Seq after water deprivation. The functional enrichment of induced transcripts revealed a significant transcriptional remodelling of perception, signalling, transcription and metabolism related to stress. Simultaneously, the induction of the enzyme 9-cis-expoxycarotenoid dioxygenase (NCED) transcripts reflects the central role of the hormone abscisic acid in this response. Consequently, the total content of amino acids and soluble sugars increased, and that of starch decreased. Likewise, to preserve cell membranes and water uptake, osmotic adjustment and radical growth were significantly promoted. These results provide a comprehensive overview of how *I. paraguariensis* roots respond to dehydration. It has potential use in the genetic improvement of the species by transferring desirable traits to more sensitive genotypes to obtain more efficient crops in water use.

POSTER #57

Transcriptome dynamics of rooting zone and leaves during in vitro adventitious root formation in Eucalyptus nitens

Paula G. Ayala[1,2], Raúl M. Acevedo[1], María J. Duarte[1], Claudia V. Luna[1], Sergio A. Gonzalez[3], Máximo L. Rivarola[3], Susana N. Marcucci Poltri[3], Ana M. Gonzalez[1], Pedro A. Sansberro[1]

1 Instituto de Botánica del Nordeste (IBONE), UNNE-CONICET, Corrientes, Argentina. **2** Present address: Mejoramiento Genético forestal. INTA-EEA Concordia, Argentina. **3** Instituto de Agrobiotecnología y Biología Molecular (IABIMO), INTA-CONICET, Hurlingham, Argentina.

The properties of wood and the agronomic attributes associated with its fast-growing and tolerance to frost turn around *Eucalyptus nitens* as a valuable forest alternative. However, the rapid maturation-related decline of adventitious root (AR) formation limits the success of vegetative propagation of selected adult trees. In order to elucidate the molecular mechanisms involved in AR formation, we set out to analyze the variation in gene expression during the early stage of the AR induction phase of *E. nitens* microcuttings and its relationship with the loss of maturation-related decline of AR formation. Through histological techniques, the ontogenic phases that comprise the developmental process were delimited, and within the quantification of the *scarecrow* and *shortroot* genes, the appropriate developmental stage of sample collection for the massive sequencing (RNAseq) was established. The analysis of the variation in the transcriptomic profiles of leaves and stems made it possible to clarify the auxin metabolism that occurs in the propagule in response to IBA treatment. Likewise, it made it possible to interpret the hormonal interaction and the signaling process leading to the formation of the root meristem. The quantification of a group of genes involved in the morphogenic process confirmed that cytokinin signaling as negative regulators of the morphogenic process determines, in part, the decrease in rooting capacity as a function of age. The analysis of the transcriptomic variation in the leaves and stem base provided profuse information that allowed us: 1) to elucidate the metabolism of auxins, 2) to understand the hormonal interaction and the signaling process involved, and 3) to collect data associated with their recalcitrance.

POSTER #59

Exploration of Focal Adhesion Kinase inhibition space with data science methods for the development of potential novel inhibitors

Quispe Patricia A.[1], Martinez H. Leandro[2], León Ignacio E.[1], Lavecchia Martín J.[1]

1 CEQUINOR (CONICET-CTT La Plata, Fac. de Cs. Exactas - UNLP). **2** INIFTA (CONICET-CCT La Plata, Fac. de Cs. Exactas - UNLP)

Focal adhesion kinase (FAK) is a 125-kDa tyrosine kinase that plays a key role in tumor adhesion, survival, motility, and angiogenesis, leading to invasion and metastasis. Several kinase domain binding inhibitors have been designed, using both experimental and computational information. Most of them bind to the ATP binding site and disrupt a salt bridge formation, reducing the kinase activity of FAK. To date, eight ATP competitive inhibitors have been tested in clinical trials, and most potential inhibitors are structural analogs of these compounds, therefore sharing similar features. Although this strategy has been very successful, it has shortcomings when it comes to the discovery of novel scaffolds. The binding information of these active compounds could improve virtual screening campaigns, opening up the chemical space of inhibitors. On other hand, databases such as ChEMBL, Pubchem, and BindingDB contain compounds with reported inhibitory activities on FAK that could exhibit a different binding mode than that of classical inhibitors. With this in mind, a detailed study of binding modes of active compounds was carried out using docking, molecular dynamics simulations, and MM/GBSA decomposition. The generated interaction profiles were then treated with data science methods in order to cluster and identify interaction patterns. Our work focuses on guiding candidate selection with active compounds so that the candidates bind in a similar fashion, but the molecular structure is not necessarily the same. Particularly in FAK, we strive to add new inhibiting scaffolds, focusing on understudied binding modes.

POSTER #61

AlphaFold2 error estimation as a function of sequence divergence

Cristian Guisande Donadio[1], Nicolas Palopoli[1], Maria Silvina Fornasari[1] and Gustavo Parisi[1]

1 Universidad Nacional de Quilmes

Background: In recent years, the computational tool AlphaFold2 (AF) achieved an impressive performance in predicting 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 relative distances between residues. Those representations are further improved using an end-to-end approach to generate structure models with iterative structural refinement. Besides these outstanding achievements, AF fails to predict single mutation effects on protein structure. As AF extracts and derives structural information from MSAs, it is possible to deduce that the same MSA will be used for very close sequences and then it will be difficult to model their structural differences. In this work we explore the error of AF predictions as a function of sequence divergence.

Results: To answer this question, we used Uniprot proteins with known structural information inferred from X-ray crystallography (PDB structures). Our dataset is composed of 3762 pairs of PDBs and their corresponding AF models. We derived an error score based on the structural deviation between known PDB and AF models as a function of the sequence divergence between the corresponding pairs of proteins being compared. We found that the highest errors are found in proteins with more than 90% of sequence identity. Although the error is still high up to about 70% identity, below that level of divergence the error starts to be the same even between very dissimilar sequences (below 30%). Also we found that the error correlates with sequence divergence only within the range of 90-100% sequence identity.

Conclusions: We think that the estimation of the error as a function of the sequence divergence could help in a better application of this extraordinary tool in evolutionary studies, where a significant sequence divergence is expected.

POSTER #62

Expanding the repertoire of human tandem repeat RNA-binding proteins

Agustín Ormazábal[1,2,3], Matías Sebastián Carletti[2,3], Tadeo Enrique Saldaño[2], Martín Gonzalez-Buitron[2,3], Julia Marchetti[2], Nicolas Palopoli[2,3], Alex Bateman[1]

1 European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton. CB10 1SD. UK. **2** Departamento de Ciencia y Tecnología, Universidad Nacional de Quilmes, Bernal, Buenos Aires, Argentina. **3** CONICET, Buenos Aires, Argentina.

Background: Protein regions consisting of arrays of tandem repeats are known to bind other molecular partners, including nucleic acid molecules. Although the interactions between repeat proteins and DNA are already widely explored, studies characterising tandem repeat RNA-binding proteins are lacking. We performed a large-scale analysis of human proteins devoted to expanding the knowledge about tandem repeat proteins experimentally reported as RNA-binding molecules. This work is timely because of the release of a full set of accurate structural models for the human proteome amenable to repeat detection using structural methods.

Results: We identified 219 tandem repeat proteins that bind RNA molecules and characterised the overlap between repeat regions and RNA-binding regions as a first step towards assessing their functional relationship. Our results showed that the combination of sequence and structural methods finds more tandem repeat proteins than either method alone. We observed differences in the characteristics of regions predicted as repetitive by sequence-based or structure-based computational methods in terms of their sequence composition, their functions and their protein domains.

Conclusions: We have carried out a comprehensive screening for RNA-binding proteins with tandem repeats and found that the fullest coverage of repeats is achieved when using both sequence and structure-based methods, which can be viewed as complementary in nature. Our results also emphasise the complementary roles of both ordered domains and tandem arrays as well as disordered regions in the RNA-binding process.

POSTER #64

Characterization of short linear motif-mediated interactions in amyloid proteins

Juan Griffin[1], Juan Mac Donagh[1,2], Gustavo Parisi[1,2], Nicolás Palopoli[1,2]

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

Background: Amyloidogenic proteins undergo extreme conformational changes under certain conditions, usually related to disease. Many tend to be enriched in intrinsically disordered regions which usually harbor Short Linear Motifs (SLiMs) mediating protein interactions. Although the role of SLiMs in some amyloidogenic proteins has been explored, a wider analysis is missing. Here we seek to explore the presence of SLiMs related with amyloid proteins. We rely on motif discovery methods and computational predictions based on sequence conservation, along with manual curation from literature. As our initial dataset we explore reports of SLiM interactions in biofilms and Alzheimer's disease. The following step involves the identification of SLiMs in a large and diverse set of amyloids previously obtained in our group.

Results: The CsgA protein from *E. coli* integrates the curli amyloid fibrils of biofilms. CsgA carries the known motif 22-VVPQYG-27 that facilitates its interaction with CsgG. This motif does not match any known class in ELM, the reference database of SLiMs in eukaryotes and their pathogens. We were able to retrieve and align 33 homologous bacterial CsgA sequences and used SLiMMaker to identify the expression [ASV][IV]PQ[FGWY]G as its conserved pattern. The human amyloid precusor protein (APP) generates amyloid-beta polypeptides associated with Alzheimer's disease. 682-YENPTY-687 is a known instance of the ELM class LIG_PTB_Apo_2. Its last position has been reported as part of another motif 687-YKFFE-691, responsible for an interaction with AP4 that we found conserved in a deep alignment of animals. We also observed that a third motif, 653-YTSI-656, that is phosphorylated to block the interaction of an adapter protein that redirects to lysosomes, is not listed in ELM. However, it matches three class definitions, including TRG_ENDOCYTIC_2, with the role of binding adapter proteins like AP4.

Conclusions: We propose a novel motif class responsible for transport of CsgA in the curli biogenesis system. We also analyzed the APP human protein and identified two conserved motifs that may constitute new SLiM classes. Our preliminary results support the feasibility of this approach to analyze the presence of SLiM-mediated interactions in a larger dataset of 81 metastable amyloidogenic human proteins.

POSTER #66

Data analysis of food bioactive peptide for the design and construction of a new database

Manuel Lonigro[1], Virginia González[2], Sebastián Bassi[2], Nicolás Palopoli[1,3], Agustina Nardo[1,4,5]

1 Departamento de Ciencia y Tecnología, Universidad Nacional de Quilmes, Bernal, Bs. As, Argentina. **2** Toyoko LLC, Berkeley, California, EE. UU. **3** Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Argentina. **4** Laboratorio de Investigación, Desarrollo e Innovación en Proteínas Alimentarias, CIDCA, La Plata, Bs. As, Argentina. **5** Departamento de Ciencias Biológicas, Facultad de Cs. Exactas – UNLP, La Plata, Bs. As, Argentina.

Background: Bioactive peptides (BPs) derived from food proteins are sequences encrypted in proteins of different origin that, after their release *in vitro* and/or *in vivo*, are able to exert one or more physiological effects (activities) in humans, depending on their physicochemical and structural characteristics. The characterization and identification of BPs is of great interest for both the food and pharmaceutical industries. There is a need to systematize the available information on bioactive peptides derived from food proteins and in particular to develop a versatile database to optimize the study and characterization of new BPs in order to facilitate quantitative structure-activity relationship analyses. Here we explore existing databases of BPs as a first step towards the integration and extension of available knowledge.

Results: A survey of BPs databases was performed. Those with information on BPs derived from food proteins were selected: BIOPEP-UWM, PlantPepDB and FermFooDb. We analyzed the number of entries (peptides) recorded, the biological activities in which they are grouped, activity parameters (IC 50 /EC 50), length distribution and amino acid composition of all records and by activity. The records were compared and the initial dataset was constructed. This was extended with structural information available in the PDB and AlphaFold database, and obtained *ad hoc* by comprehensive molecular dynamics simulations of tri - and penta-peptide structures.
(https://registry.opendata.aws/short_peptides/).

Conclusions: Our work allowed us to retrieve data to design and build a novel database of bioactive peptides with functional and structural annotations. It will be made available through a public web page with data that will be of academic and industrial utility.

POSTER #67

Conformational characteristics of LTSV40 intrinsically disordered regions and their implication in its pRb binding mechanism

Carla Luciana Padilla Franzotti[1,2], Nicolas Palopoli[1,2], Gustavo Pierdominici-Sottile[1,2]

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

Background: Human pocket proteins, such as retinoblastoma (pRb), are negative regulators of the eukaryotic cell cycle. They repress the expression of certain genes and are thus considered tumor suppressors. The large T antigen of Simian Virus 40 (LTSV40) is a multifunctional protein with two intrinsically disordered sequences that links three distinct domains. One of these disordered regions includes a pentapeptide linear motif (SLiM) that binds pRb causing cell cycle modification. In this work, by means of computational studies, we analyzed the pRb-LTSV40 binding mechanism and examined the main characteristics of LTSV40 disorder sequences before the binding event.

Results: By means of the Umbrella Sampling technique, we investigated the interaction mechanism that involves a disorder sequence of LTSV40 and the pRb binding pocket. We could identify that, during the binding process, the first recognition is due to the SLiM located within the LTSV40 disorder region. It could be also observed that part of the N-terminal flanking sequence of the SLiM adopts an alpha-helix structure upon binding. The analysis of the conformational features of the two intrinsically disordered sequences of LTSV40 in its free form showed that the number of residues in the sequence has a significant effect on the distribution of lengths that it can reach. Besides, the analysis of the conformational features of the two intrinsically disordered sequences of LTSV40 showed that this N-terminal flanking portion can be either unstructured or in an alpha-helix form. Finally, the estimation of the percentage of availability of distinct regions of the two intrinsically disordered sequences indicated that portions involved in recognition events (such as the SLiM) presents a marked solvent exposition, higher than their counterparts. Hypotheses of the structural reasons behind these facts are still under analysis.

Conclusions: The results provide valuable information about molecular mechanisms involved in the interaction between LTSV40 and the human cellular pocket protein pRb. In addition, this could contribute to the general knowledge of protein-protein interaction mediated by intrinsically disordered regions.

POSTER #68

Identification of immune systems with potential biotechnological application in bacteria of the genus Acinetobacter

Haydé Saracho[1], Virginia Marcelino[1], Daniel German Kurth[1]

1 Planta Piloto de Procesos Microbiológicos Industriales (PROIMI - CONICET)

Background: Bacteria and archaea are constantly exposed to the rampant invasion of phages or plasmids, generically known as mobile genetic elements (MGEs). In response, they have evolved diverse, complex defense systems that limit the intrusion of these foreign elements. Many new types of defense systems have recently been discovered by studying the genomic 'dark matter' of defense islands using a blame-by-association approach: uncharacterized genes that commonly reside alongside genes from known phage defense systems. they often encode new defense systems. As more genomic data is deposited in sequence databases, so are renewed efforts to comprehensively identify and characterize known defense systems. We propose to evaluate the heterogeneity of two recently developed tools: Prokaryotic Antiviral Defense LOCator (PADLOC) and DefenseFinder of in silico prediction of phage defense systems from NCBI genomes.

Results: From a Unix command line we used the rsync program and downloaded from RefSeq release 212, 8468 genomes of bacteria of the genus Acinetobacter. We make the predictions locally and compare them with each other. PADLOC predicted 121,139 defense systems of 108 different types. DefenseFinder predicted 43288 defense systems of 105 different subtypes.

Conclusions: The analyzed tools have different criteria of type and subtype of defense system. PADLOC and DefenseFinder show strong agreement and heterogeneity is not pronounced between the tools used. Most of the predicted systems were not tool-specific. We believe that the integration of PADLOC and DefenseFinder is necessary since both tools are valuable for the identification and future characterization of defense systems against phages in bacteria of the genus Acinetobacter. No single best approach was identified. The definition of the “defense islands” on each genome will provide new regions to identify potentially new systems.

POSTER #69

Identification of Koch bacili through machine learning

Jarvis Raraz-Vidal[1], Omar Raraz-Vidal[2]

1 Resident Physician of Clinical Pathology, Perú. **2** Resident Physician of Internal Medicine, Arzobispo Loayza National Hospital, Perú.

Background: Tuberculosis is a disease with a great impact on public health, so it is important to have new tools to help diagnose, reduce exposure and be economical. So artificial intelligence is a useful tool to help in the process of reading bacilloscopies. Therefore, the aim of the study was to develop and evaluate the clinical efficacy of the automated method based on artificial intelligence to identify koch bacilli in Ziehl-Neelsen (ZN) stained sheets. A pilot study was carried out. An automated method (based on AI) for the identification of mycobacteria was developed. We prepared a training data set with 85 positive and 85 negative slides with the same size and color, from ZN-stained slides scanned and published on the internet. Which were confirmed by 2 clinical pathologists confirming positive and negative lamina. A neural network model based on machine learning algorithms was created to identify Koch's bacillus through its characteristics, in addition to training the neural network to improve identification. There was a sample of 44 slides (22 positives). Sensitivity, specificity, PPV, NPV, LR were estimated.

Results: The pilot study was performed with 44 images (22 images with Koch's bacilli and 22 images without Koch's bacilli). We compared the pathologists' results obtained by separately evaluating the images and the results obtained through the neural network. The test obtained a sensitivity of 90% and a specificity of 80% by the AI-assisted method, for the detection of AFB. The PPV 82%, NPV 89%, FP 20%, FN 10%. The positive Likelihood ratio (LR) obtained 4.5 and the negative LR 0.12.

Conclusion: The artificial intelligence program presented good sensitivity and specificity to identify koch bacilli.

Asociación Argentina de Bioinformática
y Biología Computacional

a2b2c

bioinformatica_ar