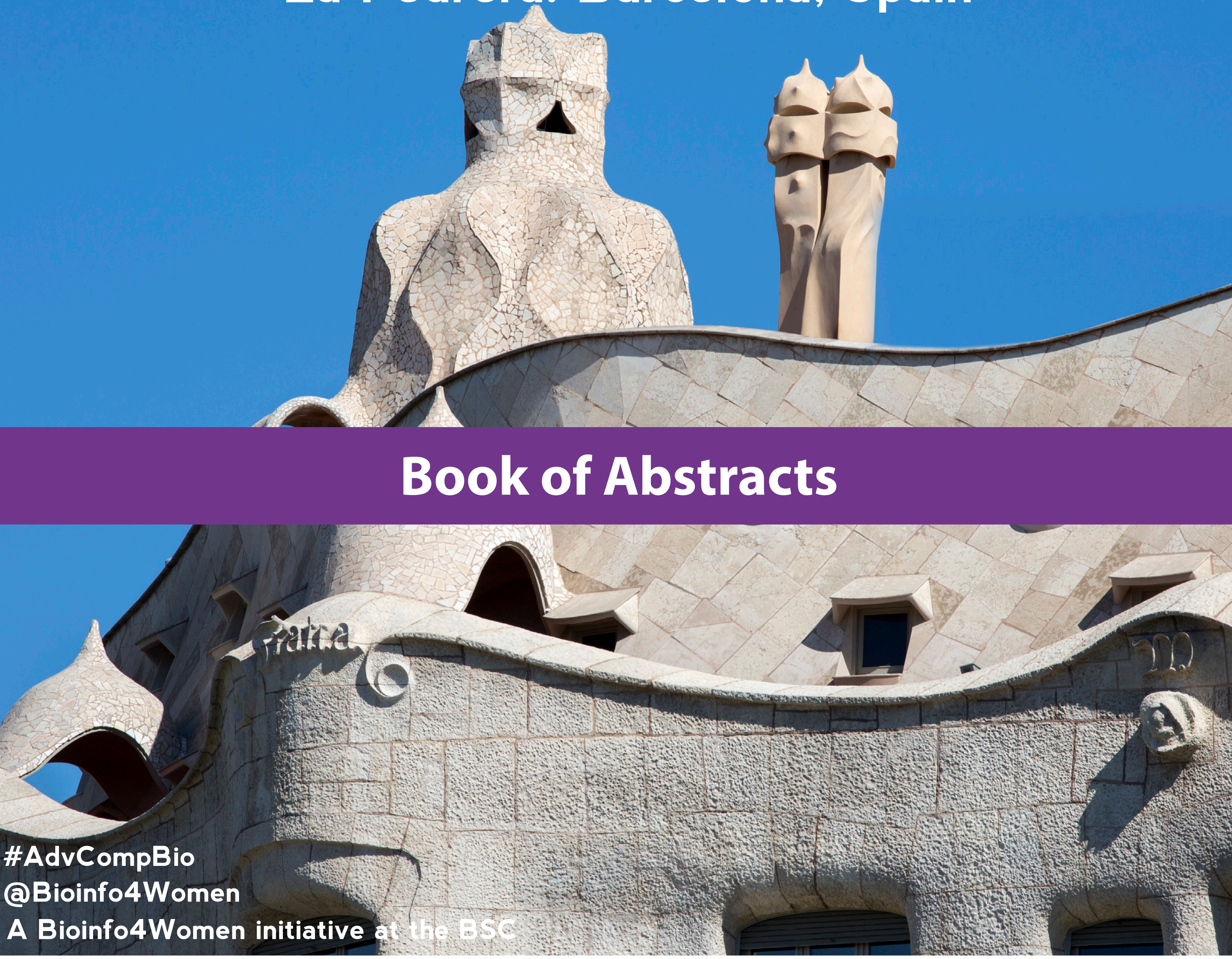
ADVANCES IN COMPUTATIONAL BIOLOGY

FOSTERING COLLABORATION AMONG WOMEN SCIENTISTS

ISCB INTERNATIONAL SOCIETY FOR COMPUTATIONAL BIOLOGY

BSC Barcelona Supercomputing Center
Centro Nacional de Supercomputación

November 28-29, 2019
La Pedrera. Barcelona, Spain



Book of Abstracts

#AdvCompBio
@Bioinfo4Women
A Bioinfo4Women initiative at the BSC

FEDER Fondo Europeo de Desarrollo Regional
UNIÓN EUROPEA "Una manera de hacer Europa"
GOBIERNO DE ESPAÑA MINISTERIO DE CIENCIA, INNOVACIÓN Y UNIVERSIDADES
Instituto de Salud Carlos III
UPC UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH
upf. Universitat Pompeu Fabra Barcelona
IMIM Institut Hospital del Mar d'Investigacions Mèdiques
Parc de Salut MAR Barcelona
INB Spanish National Bioinformatics Institute
elixir SPAIN

INDEX

PRESENTATION

We are delighted to present the first Advances in Computational Biology (AdvCompBio) conference, an initiative to promote the research done by women scientists in the fields of computation and biomedicine. In this conference all speakers and organizers are women; we aim at providing a unique opportunity to share research and set up new collaborations.

AdvCompBio is a high-level international scientific meeting, intended to discuss the latest applications of informatics to understand life systems and disease. The conference will cover a wide range of subjects including systems biology approaches, analysis of large scale genomics data and methods to evaluate drug-target interactions, among others, presented as both oral talks as posters.

By organizing this conference, we also wish to raise awareness of the existing gender bias in science and technology.

Attributes such as scientific creativity or leadership are rarely associated with women in our societies, contributing to the glass ceiling effect. The conference will showcase research by women at different stages of their careers, including prominent women scientists and directives that can serve as role models for future generations.

AdvCompBio is organized by the Bioinfo4Women programme from the Barcelona Supercomputing Center (BSC-CNS), as well as members of the Research Program on Biomedical Informatics at Hospital del Mar Research Institute (IMIM) and Universitat Pompeu Fabra (UPF), the Universitat Politècnica de Catalunya (UPC) and the Spanish National Bioinformatics Institute (INB/ELIXIR-ES). It is an affiliate conference of the International Society for Computational Biology (ISCB).

Enjoy the conference!

ORGANISING COMMITTEE

Mar Albà

Eva Alloza

Rosa Maria Badia

Nataly Buslón

Osnat Hakimi

Alba Jené

Eva Navarrete

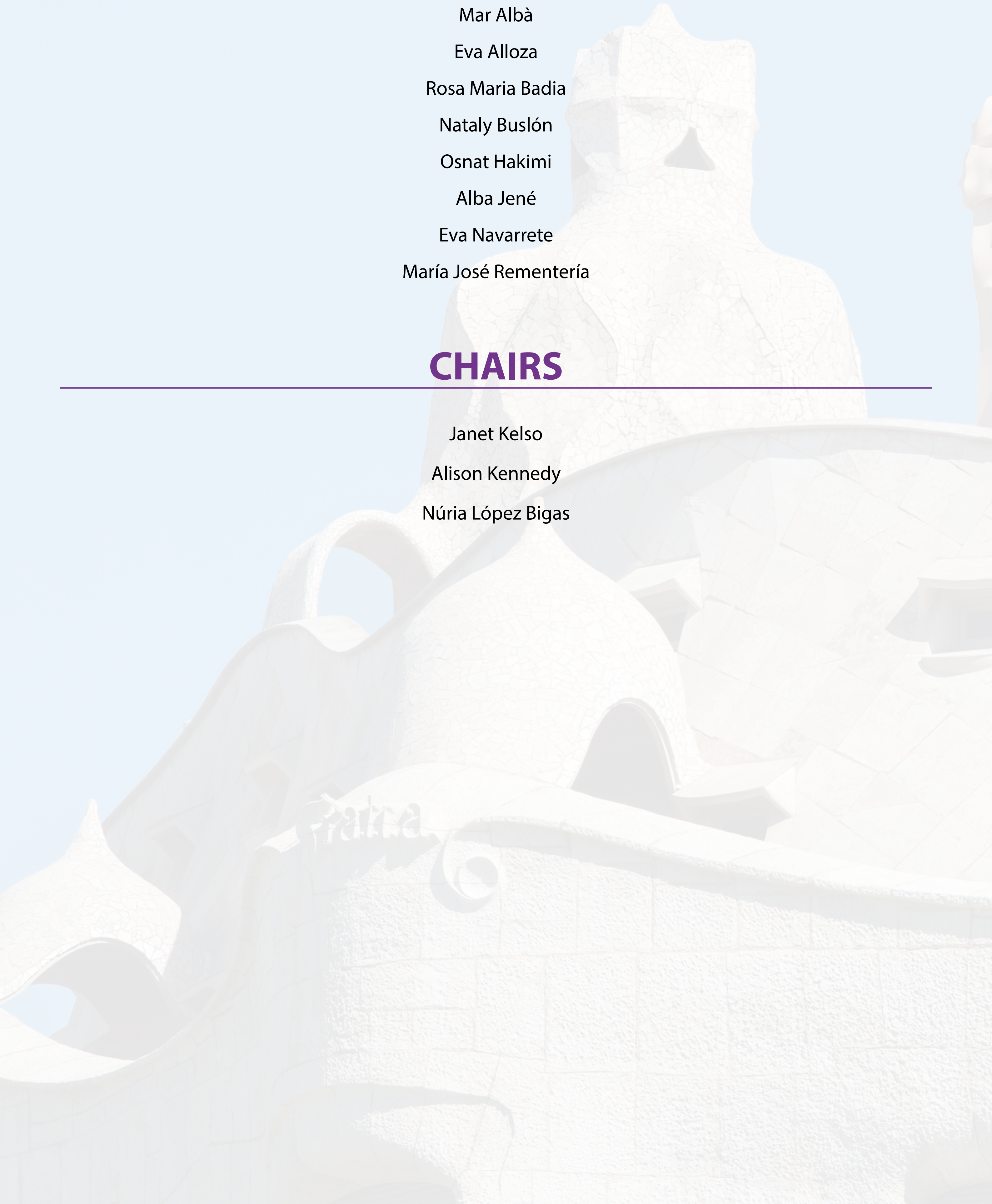
María José Rementería

CHAIRS

Janet Kelso

Alison Kennedy

Núria López Bigas



SCIENTIFIC COMMITTEE

LEARNING FROM BIOLOGICAL SEQUENCES

Ana Rojas

Marta Melé

Cristina Marino-Buslje

Eva Novoa

Anjali Gupta Hinch

Marina Marcet-Houben

Cinta Pegueroles

MACHINES SPEEDING UP RESEARCH

Fiona Reid

Michele Weiland

Dawn Geatches

Laura Schulz

WHEN COMPUTATIONAL BIOLOGY MEETS MEDICINE

Milana Frenkel-Morgenstein

Vera Pancaldi

Laura Inés Furlong

Valentina Boeva

Fatima Al-Shahrour

Anaïs Baudot

SPONSORS

Fundació Catalunya La Pedrera

FUJITSU

hp

DR. ANTONI ESTEVE FUNDACIÓ

For Women in Science

UNESCO

FONDATION L'ORÉAL

eLIFE

GCAT TACG GCAT genes

an Open Access Journal by MDPI

CAMBRIDGE UNIVERSITY PRESS

אוניברסיטת בר-אילן

Bar-Ilan University

SIB

Swiss Institute of Bioinformatics

PROGRAMME

Thursday, 28th November 2019

8:00 - 9:00	**REGISTRATION**
9:00 - 9:25	**WELCOME** – *Chair: Rosa M. Badia* -**Janet Kelso**, Group leader, Max Planck Institute for Evolutionary Anthropology -**Alison Kennedy**, Director, STFC Hartree Centre -**Núria López Bigas**, ICREA Research Professor and Group leader, Institute for Research in Biomedicine Barcelona -**Carla Conejo**, Scientific Projects Leader, Area of Knowledge, Education and Research at Fundació Catalunya La Pedrera -**Àngels Chacón**, Minister of Business and Knowledge, Government of Catalonia
9:25 - 10:10	**KEYNOTE TALK** - *Chair: Janet Kelso* **Christine Orengo (UCL). CATH Functional families - insights into impacts of genetic variations.**
10:10 - 11:10	**PLENARY SESSION 1** – *Chair: Janet Kelso* -**Compute02**: An RPCA (Robust Principal Component Analysis) based approach for protein-protein interaction hot-spot prediction, **Divya Sitani** -**BioMed03**: Predicting synthetic lethal interactions using conserved patterns in protein interaction networks, **Frances Pearl** -**BioSeq01**: Characterization of Selenoprotein Gene Expression across Tissues and Individuals, **Aida Ripoll-Cladellas**
11:10 - 11:40	Coffee break
11:40 - 13.00	**PLENARY SESSION 2** – *Chair: Alison Kennedy* -**BioSeq06**: Go low with ATLAS: a tool for maximizing insight from minimal sequencing depth, **Vivian Link** -**BioSeq02**: Tools for transforming multiomics data into disease models, **Ana Conesa** -**BioSeq05**: Genomic based drug repurposing screen for Rett syndrome, **Irene Unterman** -**Compute05**: DLMF: Deep Logistic Matrix Factorization with multiple information integration for drug-target interaction prediction, **Sarra Itidal Abbou**
13:00 - 14:00	Lunch + **Poster session I**
14:00 - 14:45	**KEYNOTE TALK** - *Chair: Janet Kelso* **Marie-Christine Sawley (Intel). Data centric large scale computing, a powerful scientific instrument for life and science.**
14:45 - 16:05	**PLENARY SESSIONS 3** – *Chair: Rosa M. Badia* -**BioSeq03**: Widespread sexual dimorphism in genetic architecture in UK Biobank, **Elena Bernabeu** -**BioSeq04**: Fine-mapping UK Biobank traits GWAS using bayesian algorithms and chromatin annotation data, **Erola Pairo-Castineira** -**Compute03**: Automated extraction of color pattern and anatomical characteristics in dairy cows, **Jessica Nye** -**Compute04**: The human heart seen by the eyes of a computer scientist, **Marta Garcia-Gasulla**
16:05 - 16:10	***Group Picture***
16:10 - 17:10	Coffee Break + ***Meeting with Women Leaders*** (prior confirmed registration) + **Poster session I**
17:10 - 18:50	**PLENARY SESSION 4** – *Chair: Núria López-Bigas* -**BioMed05**: Insight into genetic predisposition to chronic lymphocytic leukemia from integrative epigenomics, **Renée Beekman** -**BioMed01**: OncodriveCLUSTL: a sequence-based clustering method to identify cancer drivers, **Claudia Arnedo-Pac** -**BioMed02**: Detecting aberrant integrations of viral DNA that promote major restructuring of cancer genome architecture, **Eva G Alvarez** -**Compute01**: Discriminating Early- and Late-Stage Cancers Using Multiple Kernel Learning on Gene Sets, **Arezou Rahimi** -**BioMed04**: Mechanistic approach for optimal model selection in cancer research, **Inmaculada Álamo-Álvarez**
18:50 - 19:20	Theatre Performance
20:30	***Networking Conference Dinner*** (prior registration, 7 Portes Restaurant)

PROGRAMME

Friday, 29th November 2019

9:00 - 9:45	**KEYNOTE TALK** – *Chair: Núria López-Bigas* **Natasa Przulj (BSC, ICREA). Towards Data-Integrated Medicine.**
9:45 - 10:45	**PLENARY SESSION 5** – *Chair: Mar Albà* **-BioSeq26**: Dynamic hyper editing underlines temperature adaptation in Drosophila, **Ilana Buchumenski** **-BioSeq29**: Benchmarking coevolution methods, **Rocío Rama Ballesteros** **-BioMed21**: Computational challenges associated to plasmid mediated AMR spread, **Alice Ledda**
10:45 - 11:15	Coffee Break
11:15 - 12:35	**PLENARY SESSION 6** – *Chair: Marta Melé* **-BioMed22**: Cell-intrinsic core-regulatory circuits driving tumor-related phenotypes with the I3-OncoNet cycle, **Julia Puig** **-BioSeq28**: Single-cell transcriptomics analysis reveals the dynamics of alternative polyadenylation during cell cycle progression, **Mireya Plass** **-Compute15**: An atomistic molecular dynamics simulations approach to the study of h-LDHA inhibition, **Antonia Vyrkou** **-BioSeq27**: Comparative epigenomics determines the enhancer sequence code of pluripotent embryonic stem cells and facilitates the creation of synthetic enhancers, **Jennifer Mitchell**
12:35 - 14:45	Lunch + **Poster session II**
14:45 - 15:45	**Panel discussion:** Challenges of Artificial Intelligence in Biomedical Research (ACM-WE & RSG-Spain)
15:45 - 16:10	**Closing remarks and Best Poster and Oral presentation awards** – *Chairs: Janet Kelso, Alison Kennedy and Núria López-Bigas*
16:10 - 17:00	Guided visit to La Pedrera (*prior registration*)
16:10 - 17:30	***Networking activity*** (RSG-Spain)

KEYNOTE SPEAKERS

Christine Orengo, PhD. University College London (UCL)

Christine Orengo is a computational biologist, whose core research has been the development of robust algorithms to capture relationships between protein structures, sequences and functions. She has built one of the most comprehensive protein classifications, CATH, used worldwide by tens of thousands of biologists, and central to many pioneering structural and evolutionary studies.

CATH structural and functional data for hundreds of millions of proteins has enabled studies that revealed essential universal proteins and their biological roles, and extended characterisation of biological systems implicated in disease e.g. in cell division, cancer and ageing. CATH functional sites have revealed protein residues implicated in enzyme efficiency and bacterial antibiotic resistance. This data also identified genetic variations likely to be driving human diseases and the drugs that can be repurposed to offset the pathogenic effects.

Christine is a Vice President of the International Society of Computational Biology (ISCB). She is a Fellow of the Royal Society of Biology and Elected member of EMBO since 2014, and a Fellow of ISCB since 2016. She is a founder of ELIXIR 3DBioInfo.

Natasa Przulj, PhD. Barcelona Supercomputing Center (BSC-CNS)

Prof. Przulj is recognized for initiating extraction of biomedical knowledge from the wiring patterns (topology, structure) of large real-world molecular (omics) networks. She designs new algorithms for mining and integrating the wiring patterns of systems-level, heterogeneous omics networks. She applies them to uncover new biological and medical information and gain deeper biomedical understanding.

She is an elected academician of The Academy of Europe, Academia Europaea, The Serbian Royal Academy, and a Fellow of the British Computer Society. She is an ICREA Research Professor at Barcelona Supercomputing Center. She has been a Professor of Biomedical Data Science at University College London (UCL) Computer Science Department since 2016. She received two prestigious European Research Council (ERC) grants, the ERC Consolidator grant titled "Integrated Connectedness for a New Representation of Biology" (2018-2023) and the ERC Starting Independent Researcher Grant titled "Biological Network Topology Complements Genome as a Source of Biological Information" (2012-2017). She was awarded the British Computer Society Roger Needham Award in 2014 for a distinguished research contribution in computer science by a UK based researcher within ten years of their PhD. She held a prestigious NSF CAREER Award for the project titled "Tools for Analyzing, Modeling, and Comparing Protein-Protein Interaction Networks" in 2007-2011 at University of California Irvine.

Prof. Przulj is a member of the Editorial Boards of Bioinformatics (Oxford Journals), Scientific Reports (NaturePublishing Group) and Frontiers in Genetics (Frontiers), an Associate Editor of the Journal of Complex Networks (Oxford Academic) and BMC Bioinformatics (BioMed Central), a member of the Scientific Advisory Board of the Helmholtz Centre for Infection Research (HZI / Braunschweig, Germany), and the Proceedings / Area Chair / COSI Lead of Protein Interactions, Molecular Networks and Network Biology tracks at the ISMB/ECCB 2015, ISMB 2016 and ISMB/ECCB 2017, ISMB 2018 and ISMB/ECCB 2019.

Marie-Christine Sawley, PhD. Intel

Marie-Christine joined Intel France in December 2010 as the Exascale Lab Director in Paris, and since then, of the collaboration with Barcelona Supercomputing Center.

She completed a PhD in plasma physics at CRPP-EPFL in 1985 on analytical and numerical studies of the coupling phenomena between electromagnetic waves and magnetically confined plasmas. In 1997, she joined the “Mastering the Technology Enterprise” curriculum at IMD in Lausanne.

Active in HPC and scientific computing for the last two decades, Marie-Christine has been the director of CSCS, the Swiss national supercomputing centre for five years 2003 - 2008, before joining CMS Computing at CERN. Prior to this, she was one of the proponents and drivers for establishing the Vital-IT centre of the Swiss Institute for Bioinformatics in 2002 in partnership with HP and Intel.

Marie-Christine has also conducted a number of prominent activities in the area of scientific and technology, such as for the EPFL CRAY collaboration in 1994-95, EPFL Alinghi project in 2002-03, Swiss TX cluster at EPFL and promotion of ETH Zurich and Swiss industries in the building of the CMS detector at CERN in 2008.

KEYNOTE TALKS

ABSTRACTS

KEYNOTE TALKS

CATH Functional families - insights into impacts of genetic variations

Christine Orengo[1]

[1]-Professor of Bioinformatics, University College London

Powerful tools for comparing protein structures and protein sequences have allowed us to analyse proteins from more than 20,000 completed genomes and identify 5500 evolutionary domain superfamilies, comprising a total of ~90 million domains. These superfamilies cover nearly 70% of domains from all kingdoms of life and are captured in our CATH resource. Some structural frameworks seem particularly suited to supporting different residue arrangements in the active sites and structural variations on the surfaces of the domains which can modify protein functions. Sub-classification of CATH superfamilies into functional families (FunFams) allows us to examine the structural mechanisms of function evolution in these superfamilies.

We have used the CATH-FunFams to analyse the impacts of genetic variations. For example, we observe that a particular mode of alternative splicing – Mutually Exclusive Exons (MXE) – is typically associated with variation in a small subset of residues on the surface of the protein and close to known or predicted functional sites for that protein. We analysed these effects using publicly available MXE data from 5 model organisms. Some compelling examples of MXE events in glycolytic proteins have been explored in more detail. The CATH-FunFams have also be used to determine whether genetic variations linked to human disease e.g. cancer, result in changes in residues close to functional sites, thereby modifying the functions of the proteins and affecting specific pathways and processes.

Data centric large scale computing, a powerful scientific instrument for life and science

Marie-Christine Sawley1

1-Intel Switzerland

Powerful computers have been the close companion of scientific discovery for decades. The evolution of the underlying computing technology has fueled innovation and creativity in unprecedented ways; but perhaps the most remarkable leap forward of the last 20 years is to be found in the new capacities offered to biosimulation and biological data analysis. The more recent combination of HPC and DataAnalytics scalable technology is pushing further the boundaries of discoveries and frontiers. This talk will show with concrete examples how extreme computing technology has enabled progress in computational biology, while being inspired by the rich portfolio of life science through co design activities; it will also open the perspective on how promising and innovative technologies such as quantum computing and neuromorphic processors may shape the domain during the next decades.

Towards Data-Integrated Medicine

Natasa Przulj[1]

1-Group Leader, Barcelona Supercomputing Center (BSC-CNS). ICREA Research Professor

We develop methods for extracting new biomedical knowledge from the wiring patterns of systems-level biomedical omics data. Our new methods uncover the patterns from molecular networks and the multi-scale network organization indicative of biological function, translating the information hidden in the omics data into domain-specific knowledge. We introduce a versatile data fusion (integration) framework to address key challenges in precision medicine: better patient stratification, prediction of driver genes in cancer, and re-purposing of approved drugs to particular patients and patient groups. Our new methods stem from novel network science approaches coupled with graph-regularized non-negative matrix tri-factorization machine learning methods. We utilize our new methodologies for performing other related tasks, including uncovering new cancer mechanisms and disease re-classification from modern, heterogeneous molecular level data, also inferring new Gene Ontology relationships, and aligning multiple molecular networks.

ORAL PRESENTATIONS

BioMed01

OncodriveCLUSTL: a sequence-based clustering method to identify cancer drivers

Claudia Arnedo Pac[1], Loris Mularoni[1], Ferran Muiños[1], Abel González[1], Nuria López[1]

[1]*Institute For Research In Biomedicine (IRB Barcelona), Barcelona, Spain*

One of the key objectives in oncogenomics research is the identification of the genomic alterations that drive tumor development. Cancer driver genes have been computationally detected using methods based on signals of positive selection, which are acquired during tumor evolution. These signals --recurrence, clustering and high impact of somatic mutations-- have been shown to be complementary in the detection of driver genes, thus highlighting the need of combining different up-to-date methods. However, these algorithms face the challenge of accurately calculating the expected mutation rates to detect positive selection. Interestingly, recent work1 showed that mutation rates can be modelled locally --region wise-- using the probabilities of k-nucleotide context substitutions, avoiding genome-wide covariates and extending method's applicability to the non-coding regions of the genome. Until now, no clustering-based method using this model was available. Here we present OncodriveCLUSTL2, a new sequence-based clustering algorithm to detect significant clustering signals across genomic regions. OncodriveCLUSTL is based on a local background model derived from the tri- or penta-nucleotide context substitutions extracted from the cancer cohort under analysis and can be applied to coding and non-coding regions from any species using whole exome and whole genome sequencing data. Our method is able to identify known clusters and bona-fide cancer drivers in coding regions, outperforming the existing OncodriveCLUST and complementing other methods based on different signals of positive selection. OncodriveCLUSTL also highlights different non-coding regions with significant clustering signals for further characterization.

1Mularoni,L. et al. (2016) OncodriveFML: A general framework to identify coding and non-coding regions with cancer driver mutations. Genome Biol., 17, 128.

2Arnedo-Pac,C. et al. (2019) OncodriveCLUSTL: a sequence-based clustering method to identify cancer drivers. Bioinformatics., Epub Jun 22.

BioMed02

Detecting aberrant integrations of viral DNA that promote major restructuring of cancer genome architecture

Eva G Alvarez[1], Javier Temes[1], Jose Mc Tubo[1]

[1]CIMUS, University of Santiago de Compostela, Spain

Cancer is the most common genetic disease but, surprisingly, it has been estimated that viruses are the cause of about 10-15% of all cancers worldwide. There are multiple ways by which viruses can create structural variation into the human genome; one of such ways is the ‘de novo’ integration of viral DNA, a mechanism called ‘insertional mutagenesis’. During infection, full or partial sequences of the viral DNA could be integrated into the nucleus DNA of the infected cell. This is a necessary stage of the ‘life’ cycle of some viruses, such as retroviruses. However, the host nuclear genome also contains DNA from other viruses that, in principle, do not need to integrate their nucleic acid to complete their cycle.

To explore the landscape of viral DNA integration acquired somatically in cancer, we have developed a bioinformatic tool to analyse both, paired-end short reads and long reads. Short reads analysis is based on unmapped reads, discordant read pairs and the depth of coverage, while the second takes the advantages of single-molecule sequencing, looking for insertions within the long reads. To ensure that viral integrations are only present in the tumour sample, the ‘paired-mode’ of the tool compares genomes from tumoural vs normal (non-affected) tissues. However, it can be run in a ‘single-mode’, useful for cases lacking normal samples.

The paired-end analysis was used to analyze Illumina whole-genome sequencing data from 2,768 tumours and matched-normal pairs within the framework of the Pan-Cancer Analysis of whole Genomes Project. Our bioinformatic algorithms identified 266 putative somatic insertions, most of them in hepatocarcinoma, head-and-neck, and cervical tumours. From them, 74 follow the canonical patterns of a viral integration.

Apart from them, our pipeline was able to detect viral insertions following an unexpected non-canonical pattern. Interestingly, the majority (192/266) of the identified events followed this pattern, which suggest an aberrant mechanism of viral DNA integration. We had the opportunity to perform a detailed analysis of some of these abnormal events, using single-molecule sequencing technologies. This analysis revealed genomic deletions, duplications, fold-back inversions, and genomic translocations mediated by integration of viral DNA in the cancer genome. In three independent notable examples, we found that genome translocations promoted by aberrant Hepatitis B virus DNA insertions involved deletion of relevant tumour suppressor genes, including TP53 and ARID1A. These observation illuminate a relevant role of viral DNA insertion in remodeling the cancer genome in particular tumour types, representing another potential mechanism by which cancer clones acquire mutations that allow them to grow and survive.

BioMed03

Predicting synthetic lethal interactions using conserved patterns in protein interaction networks

Graeme Benstead-Hume[1], Xiangrong Chen[1], Suzanna Hopkins[2], Karen Lane[2], Jessica Downs[2], **Frances Pearl**[1]

[1]University Of Sussex, Brighton, United Kingdom, [2]The Institute of Cancer Research, United Kingdom

In response to a need for improved treatments, a number of promising novel targeted cancer therapies are being developed that exploit human synthetic lethal interactions. This is facilitating personalised medicine strategies in cancers where specific tumour suppressors have become inactivated. Mainly due to the constraints of the experimental procedures, relatively few human synthetic lethal interactions have been identified. Here we describe SLant (Synthetic Lethal analysis via Network topology), a computational systems approach to predicting human synthetic lethal interactions that works by identifying and exploiting conserved patterns in protein interaction network topology both within and across species. SLant out-performs previous attempts to classify human SSL interactions and experimental validation of the models predictions suggests it may provide useful guidance for future SSL screenings and ultimately aid targeted cancer therapy development.

BioMed04

Mechanistic approach for optimal model selection in cancer research

María Inmaculada Álamo Álvarez[1], Maria Peña-Chilet[2], Carlos Loucera[1], Joaquin Dopazo[1]

[1]Fundación Pública Andaluza Progreso Y Salud, Seville, Spain, [2]Fundacion Progreso y Salud, BieR - Centro de Investigaciones Biomédicas en Red en Enfermedades Raras (CIBERER), Spain

“Cellular and animal models are widely used to address questions in cancer research. These are often related to the search for new therapeutic targets or novel drugs that may improve current treatments and make them personalized, more efficient and effective. However, cancer models often do not fully reproduce the same conditions and functionality occurring in the actual tumor. There are limitations inherent to these models: on the one hand, non-human organisms have limited translatability to human processes. On the other hand, in xenografts or immortalized cell lines, the original conditions and characteristics might also change from those in the patient’s tissue, either because of the change in their environment or because they accumulate mutations and chromosomal aberrations. Recently, mechanistic models that link gene expression to cell functional activity are gaining popularity and have shown to be useful in modeling complex disease processes. In this work, we propose an approach based on mechanistic signaling pathways to compare the different models used for cancer research to the actual signaling landscape of the cells in patient tumors.

We collected gene expression data of patient tumors, cell lines and animal model samples from publicly available repositories (GEO, SRA, CLL, and GDC). Data were normalized accordingly and we implemented Hipathia algorithm to obtain mechanistic models from each disease model’s functional landscape. We then applied several information-based algorithms in order to learn a meaningful representation of the signaling manifold across different datasets, which allows us to modelize the transference between the signaling landscapes of the disease models and the actual patient samples. In addition, we aim to assess the suitability of each model to reproduce the disease in each group of patients.

As a result, we will create a tool that allows us to determine which cancer models are optimal to study specific pathways and metabolic processes involved in cancer. This tool will potentially make cancer research more efficient, reduce unnecessary costs and accelerate the validation procedures necessary to get therapeutic compound innovations into clinical trials.

BioMed05

Insight into genetic predisposition to chronic lymphocytic leukemia from integrative epigenomics

Renée Beekman[1], H Speedy[2], V Chapaprieta[1], G Orlando[2], P.J Law[2], D Martin-Garcia[1], J Abril-Gutierrez[3]

[1]Idibaps, Barcelona, Spain, [2]Institute of Cancer Research, United Kingdom, [3]IUOPA, Spain

Chronic lymphocytic leukemia (CLL) has a strong genetic component, evidenced by an eight-fold increased risk to develop CLL in relatives of CLL patients. Genome-wide association studies (GWAS) have provided evidence for inherited predisposition to CLL, identifying 42 genomic regions influencing CLL risk. However, efforts to define mechanisms mediating the risk at these, largely non-coding, loci have been constrained by lack of integrated genome-wide data in large CLL series. In this study, we aimed refining the gene regulatory mechanisms and biological significance of CLL risk loci.

We (i) analysed high-resolution chromatin state maps of primary CLL samples, (ii) integrated genetic, epigenetic and transcriptomic data in up to 452 primary CLL cases by quantitative trait loci (QTL) analysis, (iii) performed in silico transcription factor (TF) binding analysis using motifbreakR and (iv) studied the three-dimensional (3D) chromatin structure of normal B cells and CLL using promoter capture Hi-C data.

Eighty-one percent of the risk loci were enriched for active regulatory elements (promoters and enhancers) in CLL, suggesting a specific regulatory role for these loci in CLL pathogenesis. Additionally, at 18 risk loci we detected regulatory regions showing genotype dependent levels of genome activity (H3K27ac QTLs) and chromatin accessibility (ATAC-seq QTLs) in primary CLL cases. Moreover, within these QTLs, we defined 60 potential functional variants underlying genetic CLL predisposition. Next, we focused on underlying biological mechanisms through which genetic variants at CLL risk loci shape the regulatory genome by performing in silico TF binding analysis. We observed that genotypes associated with higher risk to develop CLL, among others, resulted in decreased binding affinity for B-cell related TFs and increased affinity for FOX, NFAT and TCF/LEF TF family members. Thus, our findings point towards a regulatory role for these TFs in CLL predisposition. Thirdly, to infer the biological significance of CLL risk loci, we identified 36 genes that showed genotype dependent gene expression levels (eQTLs) in primary CLLs. These represent potential target genes through which CLL risk loci mediate their effect and affected pathways involved in CLL pathogenesis such as immune response, wnt signalling and apoptosis. Interestingly, the eQTLs included new genes, such as TLE3, not having been associated with CLL risk before. Lastly, we observed significant 3D chromatin interactions in CLL and normal B cells between the risk loci and 15 eQTL gene loci, highly suggestive for direct regulatory links between the risk loci and expression of these genes in relation to CLL predisposition. Importantly, these analyses showed that CLL risk loci not necessarily affect expression of the nearest gene but may mediate their effect in more distant fashion, as shown for UBR5, due to long-range 3D chromatin interactions.

By (i) characterising potential functional variants that influence the risk to develop CLL, (ii) defining regulatory elements and TFs that play a role in mediating the effect of genetic variation at CLL risk loci and (iii) determining downstream genotype-dependent effects on expression of both proximal and distant target genes at these regions, we offer improved insights into the functional and biological basis of CLL predisposition.

BioMed21

Computational challenges associated to plasmid mediated AMR spread

Alice Ledda[1]

[1] *Imperial College London, United Kingdom*

AntiMicrobial Resistance (AMR) is expected to be the major cause of death by 2050 if the current trend continues. AMR spread threatens many everyday medical procedures such as surgeries or organ transplants. A specially worrying case is when AMR genes are carried on mobile genetic elements such as plasmids. Their ability to be horizontally transmitted to neighbouring bacteria makes them a special concern for public health because they can facilitate the creation of multi-resistant bacteria.

Computational approaches to reconstruct the evolution, Horizontal Gene Transfer (HGT) and spread of plasmids and associated AMR genes represent a very promising avenue to tackle the AMR epidemics and prevent disruptive consequences for public health. I will discuss some of the computational approaches and challenges related to plasmid evolution and AMR spread, both worldwide and at small epidemiological scales.

First, I will present a very accurate experimental dataset on the genetic composition of worldwide multi-resistant E. coli plasmids, encompassing their global diversity in gene composition and transmission machinery. I will highlight the difficulties posed to the reconstruction of plasmid evolutionary histories by the variability of their genetic content. I will present new computational methods based on network theory we designed to overcome these difficulties. The resulting worldwide phylogenetic patterns are indicative of extreme rates of transmission of AMR among different plasmids and bacteria.

Second, I will explain the public health risks and challenges posed by plasmid-mediated AMR outbreaks. I will focus on the challenges associated to short-term plasmid phylogenetic reconstruction and genetic epidemiology in small healthcare settings. I will show how we tackled these challenges in a plasmid-borne carbapenemase outbreak in a hospital ward. The backbone of the outbreak was sustained by a previously undescribed association of bacterial and plasmid type, with frequent HGT to other bacterial species.

Finally, I will discuss the importance of accurate phylogenetic reconstruction and the role of modern genetic epidemiology to inform public health approaches to AMR and pathogen spread, as well as the computational challenges to provide these results in real time during outbreaks.

Cell-intrinsic core-regulatory circuits driving tumor-related phenotypes with the I3-OncoNet cycle

Julia Puig[1], Aurélien Dispot[1], Wajdi Dhifli[1], Mohamed Elati[1]

[1]*Université de Lille, France*

The study of cancer through the analysis of regulatory networks (metabolic networks, signaling pathways, transcription factor regulatory networks) should allow to a better understanding of tumor progression. The knowledge of tumor networks and of their plasticity can also be used to predict drug response and identify optimal therapeutic strategies. The identification and characterization of networks are possible through both an experimental approach and a bioinformatic approach using large scale data. These two approaches are tightly linked, the results of the bioinformatic approach guiding experimental validation and the results of the experimental approach helping to adjust and validate the bioinformatics tools developed.

We propose a computational systems oncology cycle, that we call I3-OncoNet, composed of three building blocks: inference (learn and reverse engineer), interrogation (mine and analyze) and genomic implementation (design and engineer) of oncology regulatory networks. Starting from context-specific cell lines gene expression data, the inference block finds the so-called master regulators (MR) and their core-regulatory circuits underlying transcriptional programs [LICORN:2, CoRegNet:3]. In the interrogation block, a network-based regulatory signal [LatNet:4, 5] of patient transcriptomic data is computed from the information captured in the context-specific core-regulatory circuits, resulting in a dimensionality-reduced and richer patient signal: the regulatory activity. Machine learning is applied to the new signal to obtain tumor subtype-specific regulatory features. The implementation block goes back to the context-specific cell lines by assigning them to cancer subtypes through supervised classification. Finally, the subtype-specific features are used to propose in a data-driven way potential experiments for validation of master regulators on the context-specific cell lines.

We apply the I3-OncoNet cycle to normal urothelium cell lines and patient bladder cancer transcriptomic data. Our selected tissue has well-characterized experimentally tractable normal cell systems that can be manipulated to reflect different states of differentiation and phenotype [6]. Tumor cells use and modify the networks which already exist in the normal cells of origin. It is therefore interesting to be able to study in parallel the networks in tumors and in the matched normal cells. We use published data [1] and available transcriptomic and genomic data from our collaborators (Radvanyi lab for bladder cancer and Southgate lab for normal urothelial cells). Our approach generates possible networks for both the normal urothelium, and for luminal and basal tumors. Interrogation of the networks allows us to: (1) understand the similarities and differences between cancer subtypes; (2) pinpoint the differences in active networks between normal and cancer cells; (3) from functional genetic screens using ShRNA or CRISPRi-Cas9 technologies, validate several transcription factors whose known functions make them interesting candidates to be involved in basal tumors (SOX9, involved in stemness, SNAI2 in EMT, TP63 in basal epithelium formation) or luminal tumors (TEAD3, GRHL2, FOXA1, GATA3).

[1]Ghandi et al. Nature,(2019).
[2]Elati et al. Bioinformatics, 23:2407-2414,(2007).
[3]Nicolle R, Radvanyi F, Elati M. Bioinformatics, 31(18):3066 -8(2015).
[4]Dhifli W, Puig J ,Dispot A, Elati M. BMC Bioinformatics, 19(Suppl13):426,(2019).
[5]Picchetti T, Chiquet J, Elati M, Neuvial P, Nicolle R, Birmelé E. BMC Systems Biology(2015).
[6]Fishwick C, et al. Cell Death Differ. 24(5):809-818(2017).

BioSeq01

Characterization of Selenoprotein Gene Expression across Tissues and Individuals

Aida Ripoll Cladellas[1], Didac Santesmasses[2], Roderic Guigó[1]

[1]Barcelona Supercomputing Center, Barcelona, Spain, [2]Harvard University, United States

The essential trace element Selenium (Se) has fundamental importance to human health. As a constituent of the amino acid selenocysteine (Sec), Se has relevant enzymatic and structural roles in the 25 human selenoproteins, which serve diverse biological functions (e.g., antioxidant defense or redox regulation). Although studies in mice have been instrumental to understand the function and regulation of expression of numerous selenoproteins, no studies have assessed their expression in human tissues. We used more than 18 thousand RNA sequencing (RNA-Seq) samples from the Genotype-Tissue Expression (GTEx) project to carry out a large-scale survey on the expression of selenoprotein genes, Sec machinery and cysteine (Cys) homologs across multiple human tissues and individuals. We correlated their expression with the subject phenotypes and genotypes. Even though the expression of all 25 human selenoproteins was observed in the entire set of tissues, several selenoproteins showed a tissue-selective expression. We distinguished four selenoproteins with sex-differential expression (DIO2, GPX1, GPX3, and SELENOM), and three genes showed a negative correlation of expression with age in the brain (MSRB1, SELENOT, and SELENOI). We identified 6,913 candidate expression quantitative trait locus (eQTLs) in the whole set of genes analyzed here and determined a set of potential causal variants in SELENOP using colocalization analysis with genome-wide association studies (GWAS). Thus, this study reports a characterization of human selenoprotein expression at an unprecedented scale which can help in understanding the biological role of selenoproteins, providing new insights into the relevance of Se in human health.

BioSeq02

Tools for transforming multiomics data into disease models

Sonia Tarazona[3], Carlos Martinez[2], Rafael Hernandez de Diego[2], Leandro Balzano[1], **Ana Conesa[1]**

[1]University of Florida, Gainesville, United States, [2]Centro de Investigaciones Príncipe Felipe, Spain, [3]Universitat Politècnica de València, Spain

The diversity of omics technologies available to researchers has increased in the last years and with this also the possibility of adopting the multiomics approach in genomics research. However, depending of the technology, data have very different numeric nature, dimensionality, error rates and dynamic range, aspects that are frequently ignored. To fully leverage the wealth of information provided by multiomics data, specific bioinformatics software and data analysis approaches are required. Tools are needed not only for integrative analysis, but also for storage, experimental design, visualization and modeling of these complex data. In this communication I will introduce a wide arrange to tools developed within the EU STATegra project tailored to the multiomics scenario. These include STATegraEMS for the annotation of multiomics projects, MOSim for the simulation of multiomics data, MultiPower for sample size estimations in multiomics experiments, Paintomics 3 for the pathway-based visualization of multiomics data and MORE for the modeling of gene expression as a function of multiple regulatory layers. We will present the application of some of these resources to the analysis of a longitudinal multi-omics study of type 1 diabetes. We integrated transcriptomics, metabolomics and clinical data from a cohort of individuals at T1D risk, monitored from birth to the age of 15 or positive diagnose. Using a multidimensional tensor and multivariate discriminant analysis we were able to extract a ~1100 multiomics feature signature that predicts onset of autoimmunity as early as 12 months before disease diagnosis with current biomarkers. The data were used to create a mechanistic model of disease that links energy imbalance with autoimmunity. These results propose the first model of T1D disease progression within one year time frame and opens the possibility for early intervention to stop or ameliorate the development of autoimmunity.

BioSeq03

Widespread sexual dimorphism in genetic architecture in UK Biobank

Elena Bernabeu[1], Albert Tenesa[1], Oriol Canela-Xandri[1], Konrad Rawlik[1], James Prendergast[1]

[1]*The Roslin Institute, Edinburgh, United Kingdom*

Human sexual dimorphism is observed in risk, incidence, and prevalence across a wide array of diseases and complex traits, despite males and females sharing nearly identical genomes. Sex can be considered an environmental factor, providing the genome with a distinct hormone milieu, differential gene expression, and environmental pressures arising from gender societal roles. Gene by sex interactions (GxS) could then be the culprit of some of the dimorphism observed, giving rise to sex-specific genetic effects and differences in genetic architecture. The extent and basis of these interactions is yet poorly understood. Here we provide insight into both the spread and mechanism of GxS across the genome of circa 450,000 individuals of White ethnicity and 530 complex traits in UK Biobank. We report sex differences in heritability, genetic correlation, and genetic effects for a large portion of the traits studied. This study marks the largest look into the genetics of sexual dimorphism in both depth and breadth to date and warrants the need for future studies to evaluate different clinical practices between the sexes and sex-stratified analyses.

BioSeq04

Fine-mapping UK Biobank traits GWAS using bayesian algorithms and chromatin annotation data

Erola Pairo-Castineira[1], Albert Tenesa[2]

[1]Mrc - Human Genetics Unit. University Of Edinburgh, Edinburgh, United Kingdom, [2]The University of Edinburgh, United Kingdom

In recent years, Genome-wide association studies (GWAS) have identified thousands of variants related to different traits. But, advancing from the significant to the causal SNPs to understand the underlying biology, remains a challenging issue. With the popularity of CRISPR experiments, allowing the editing of single nucleotides, it is important to know the causal SNP, in which tissue has effect, and how can we see this effect, in order to design the correct experiment. Most developed approaches to find the causal SNPs include Bayesian methods which find a reduced set of credible SNPs, and the use of functional data such as Histone marks or gene expression. This approaches are usually computationally expensive and rely on the previous knowledge about relevant tissues for the trait.

Here we propose a methodology that, from GWAS summary statistics, calculates a subset of credible SNPs with a Bayesian method, selects the relevant tissues using annotation data, and finally applies machine learning techniques to find a set of causal SNPs as well as the affected tissues and marks, and we apply it to 778 traits in UK Biobank.

Summary statistics from 778 traits were obtained from Geneatlas (Canela-Xandri et al., 2018), a database of associations for ~450,000 european individuals in UKBiobank. DHS and 7 Histone marks from ~100 tissues were obtained from Roadmap epigenomics, Transcription Start Sites (TSS) for ~ 100 tissues, from FANTOM5, and missense annotations from the ENSEMBL database.

First, the Probabilistic Inference of Causal SNPs (PICS) (Fahr et al., 2015) algorithm was used to reduce the significant SNPs to a set of credible SNPs for each trait. More than 50% of the loci could be reduced to a subset <=5 credible SNPs, with and enrichment with p-val 1e-7 for missense mutations.

Parallel to this analysis, we ran GARFIELD (Iotchkova et al., 2019), to find enrichment in the overlap between significant SNPs for a trait and a tissue annotation, for the 9 histone marks in the 100 tissues. Some traits like height are enriched for almost all tissue/marks as expected while other as cornea thickness have only enrichment in the eye and some fetal tissues.

Finally, we found the overlap between the credible SNPs and marks in significant tissues to further reduce the set of SNPs, and then, machine learning techniques and UK Biobank non-european individuals were used to select the final SNPs.

Using BMI, we were able to reduce the initial set of significant SNPs to a subset of 28171 credible SNPs. From GARFIELD, we selected the tissues with enrichment in more marks: brain (8/9), ES cells (7/9) and adipose tissue (6/9), and then the SNPs overlapping marks in these tissues (final set of 1859 SNPs). Finally a LASSO model in non-europeans was used to select the causal SNPs. The SNP with strongest effect to BMI was rs34246968, a missense variant in SEC16B,a gene associated with obesity in differen populations, (Sahibdeen et al., 2018, Hotta et al., 2009) and the SNPs with a most negative association is rs1229984, related to non-consumption of alcohol.

BioSeq05

Genomic based drug repurposing screen for Rett syndrome

Irene Unterman[1], Chaya Brodie[2], Bruria Ben Zeev[3], Yuval Tabach[1]

[1]The Hebrew University, Jerusalem, Israel, [2]Bar Ilan University, Israel, [3]Sheba Medical Center, Israel

MECP2 is a key transcriptional regulator of neuronal function and development. Mutations impairing the normal functions of MECP2 are mainly associated with Rett syndrome but also involved in other neurological disorders. The mechanisms of MECP2 action are poorly understood, limiting drug research and discovery. In recent years, an exponential growth in biomedical data has led to the development of novel approaches to mapping gene networks. Characterization and identification of the gene network associated to MECP2 will allow elucidation of its mechanism of action.

Phylogenetic profiling (PP) is an unbiased approach for constructing gene networks. PP follows the evolutionary patterns of genes across the tree of life and identifies co-evolved genes. We and others have shown that co-evolved genes are functionally associated. We combined this approach with data integration to construct a network of MECP2 interacting genes and selected drug targets within the network.

Our objective is to systematically map MECP2 interacting genes, gain insight into its mode of action, predict which drugs would benefit RTT, and validate the top hits. This will allow us to repurpose drugs for Rett syndrome (RTT), using PP, computational network analysis, and in-vitro screening.

To identify the MECP2 network and map drug targets we: 1. Mapped the evolutionary pattern of MECP2 across 1000 species and identified co-evolved genes. We scored the conservation of each human gene in every species, normalized to the overall distance from that species. The genes with the highest Pearson correlation to MECP2 were investigated further. 2. Integrated multiple data sources, including drug-gene interaction databases such as DGIDB, DrugBank and Drug Target Commons, to generate candidates for repurposing. 3. Prioritized leading compounds based on their known pharmacological properties, such as blood-brain-barrier permeability, safety, and association with the MECP2 network. 4. Analyzed the top drug candidates utilizing human MECP2 knock-out cells.

In summary, mapping MECP2 interactions may help elucidate the complex role of MECP2 and reveal key downstream elements, allowing for mechanism-based drug discovery or design. This pipeline has massive potential to advance RTT research at both the basic and translational levels. Any genetic partner identified computationally will improve our understanding of the basic regulatory mechanisms disrupted in RTT. Additionally, in-vitro drug screening is faster and more economical compared to animal models. By applying state-of-the-art computational and molecular tools to comprehensively map MECP2 we hope to gain significant insights into MECP2's mode of action and find potential therapies to alleviate RTT symptoms.

BioSeq06

Go low with ATLAS: a tool for maximizing insight from minimal sequencing depth

Vivian Link[1], Athanasios Kousathanas[2], Zuzana Hofmanová[1], Carlos Reyna[1], Zoé Pochon[1], Jens Blöcher[3], Christoph Leuenberger[4], Joachim Burger[3], Daniel Wegmann[1]

[1]Department of Biology, University of Fribourg, Fribourg, Switzerland, [2]Center for Integrative Genomics, University of Lausanne, Switzerland, [3]Paleogenetics Group, University of Mainz, Germany, [4]Department of Mathematics, University of Fribourg, Switzerland

Many methods in population genomics rely on called genotypes as input. However, especially at low depth, calling genotypes is error-prone, thus uncertain genotypes are usually filtered out based on the genotype quality. But filtering causes biases, often leading to an underestimation of genetic diversity. Alternatively, the issue of genotyping uncertainty can be solved using a probabilistic approach in which hierarchical parameters (e.g. genetic diversity) are inferred by integrating over all possible genotypes at each locus.

We here present three tools, based on this philosophy, that quantify key evolutionary quantities of genetic diversity directly from sequence alignments of individuals or populations. First, we quantify heterozygosity within genomic windows under Felsenstein's 1981 substitution model. Second, we measure the pairwise genetic distance between individuals, which can also be used to infer relatedness or perform a Multidimensional Scaling Analysis without genotype calls. Third, we quantify population structure by inferring the deficit in heterozygous genotypes as measured by the inbreeding coefficient. Using simulations as well as downsampling experiments of real data, we show that all these methods perform well even at very low mean sequencing depth, often at or below 1x. As such, these methods allow to invest in more samples rather than higher depth, and hence to increase statistical power when characterizing evolutionary processes.

However, all methods based on genotype likelihoods require these to accurately reflect the genotype uncertainty. For this, base sequencing quality scores must be carefully recalibrated, for which we present a new method particularly suited for low-depth data that does not rely on reference genome data, but exploits homozygous or conserved regions in the genome.

All our tools are implemented in our well-documented and user-friendly program ATLAS, which can readily be used in combination with other tools such as ANGSD or GATK. ATLAS is particularly suited for ancient samples that have generally low endogenous DNA content and are affected by Post-Mortem Damage (PMD), a process that causes the replacements of cytosine with thymine and leads to mutations that are not reflective of a sample's diversity. While PMD is usually addressed by removing or down-weighting potentially damaged data, ATLAS explicitly accounts for PMD in the genotype likelihoods, enabling an unbiased and more powerful comparisons between ancient and modern samples.

To illustrate the power of ATLAS, we used it to infer the origin of 18 soldiers from a colossal Bronze-age battlefield in norther Germany. This battlefield, which involved thousands of warriors, challenges the view of a lack of large-scale social organization in norther Europe during that era.

BioSeq26

Dynamic hyper editing underlines temperature adaptation in Drosophila

Ilana Buchumenski[1], Erez Levanon[1]

[1]Bar-Ilan University, Ramat-gan, Israel

In Drosophila, A-to-I editing is highly prevalent in the brain and mutations in the editing enzyme dADAR correlate with specific behavioral defects. As editing sites are usually engaged in secondary structures, temperature is predicted to impact the level and nature of editing. Here we demonstrate a role for ADAR in temperature adaptation in Drosophila. Briefly, we found that despite the higher level of editing at lower temperatures, there are more editing sites at 29°C. This is due to a less specific activity of ADAR at this temperature, which edits sites which are less evolutionary conserved, more disperse, less committed in secondary structures and more likely to be located in exons. These results strongly support the notion that at 29°C, RNA editing is less deterministic and might even have deleterious effects. Interestingly, hypomorph mutants for ADAR display a weaker transcriptional response to temperature changes. In addition, and in agreement with the differences on the head transcriptome, ADAR Hypomorph flies display a highly abnormal behavioral response while adapting to temperature changes. In sum, our data shows that ADAR is essential for proper temperature adaptation, a key behavior trait, which is essential for the survival of flies in the wild.

BioSeq27

Comparative epigenomics determines the enhancer sequence code of pluripotent embryonic stem cells and facilitates the creation of synthetic enhancers

Gurdeep Singh[1], Shanelle Mullany[1], Sakthi Moorthy[1], Virlana Schuka[1], Tahmid Mehdi[2], Richard Zhang[1], Ruxiao Tian[1], Alan Moses[1], Jennifer Mitchell[1], **Jennifer Mitchell[1]**

[1]University of Toronto, Toronto, Canada, [2]University of Toronto, Canada

Mammalian genomes are 97-99% non-coding and the function of most of this DNA remains unknown. Enhancers are a major component of the non-coding genome, functioning to regulate gene expression in a tissue specific manner by binding transcription factors. Enhancers play a major role in evolution, disease and development; however, their identification is challenging due to variable location from their target gene and the non-significant increase in overall sequence conservation compared to surrounding non-coding sequences. Many of the critical transcription factors required to maintain the embryonic stem (ES) cell regulatory network are known, however, we do not currently understand how many and which specific transcription factors are required to build a functional enhancer. A computational method was devised to identify the sequence code that confers enhancer activity in ES cells by decoding regions with conserved enhancer chromatin features (CHEF: transcription factor bound regions with high histone H3 K27 acetylation) in both mouse and human ES cells. CHEF regions are highly enriched in validated enhancers and depleted in tested regions that displayed no enhancer activity. Machine learning was applied to regions bound by at least one transcription factor in mouse ES cells to determine the sequence features that distinguish CHEF regions in mouse and human from Conserved Low Enhancer Feature (CLEF) regions. Analysis of CHEF compared to CLEF regions revealed higher purifying selection pressure for transcription factor binding motifs (TFBM). To rank the importance of specific TFBM required to build an ES cell enhancer, LASSO (least absolute shrinkage and selection operator) was used to identify the TFBM most often present and conserved in CHEF regions. LASSO identified TFBM for transcription factors known to be involved in ES cell pluripotency maintenance as well as for transcription factors not yet characterized in this context. Mutagenesis of novel TFBM in a luciferase reporter vector revealed these do contribute to enhancer activity. Comparative sequence analysis in CHEF regions of 5 species revealed 14 as the average number of TFBM required for a functional enhancer, however, there is flexibility in TFBM usage. To test this flexibility, we introduced multiple TFBM into a transcription factor bound region lacking enhancer activity. Introduction of different subsets of TFBM conferred enhancer function, while loss of TFBM already present resulted in loss of gained enhancer function. Based on the features learned from natural enhancers we built synthetic enhancers containing 10-14 unique TFBM and determined in functional assays that these had activity comparable to natural enhancers. By contrast, 14 repetitions of the one most overrepresented TFBM (OCT4:SOX2 co-TFBM), or the use of only 4 unique TFBM displayed activity 5-10 fold less than the synthetic and natural enhancers revealing TFBM diversity is an important sequence feature of enhancers. Our analysis revealed a larger and flexible repertoire of TFBM used to build a functional enhancer than previously considered. Although many TFBM within conserved enhancers are conserved across species this flexible repertoire could allow for sequence evolution without loss of function that can fine tune regulatory control.

BioSeq28

Single-cell transcriptomics analysis reveals the dynamics of alternative polyadenylation during cell cycle progression

Mireya Plass[1], Daniel Schwabe[2], Samantha Praktiknjo[2], Sara Formichetti[2], Anastasiya Boltengangen[2], Jonathan Alles[2], Martin Falcke[2], Christine Kocks[2], Nikolaus Rajewsky[2]

[1]Centre For Genomic Regulation, Barcelona, Spain, [2]Max Delbrück Center for Molecular Medicine, Germany

“Single-cell transcriptomics has revolutionized the way we can study the dynamics of gene regulation and its impact in cell proliferation and differentiation. However, specific methods to quantify transcripts isoforms at the single-gene level are lacking. This prevents the exploration of dependencies that may exist across gene isoforms and a deeper understanding of how different post-transcriptional regulatory processes such as splicing and polyadenylation are coregulated.

We have developed a new method to quantify the expression levels of gene isoforms produced using insert size variation Drop-seq (isv-Drop-seq). Using this method, we have been able to quantify the thousands of individual isoforms generated by alternative polyadenylation and study their variability across cells. Additionally, we have developed another method to computationally sort cells along the cell cycle and study the dynamics of these isoforms. Our preliminary results show that many oscillating genes use specific 3’ isoforms in different cell cycle phases. Together, these results suggest that the choice of 3’UTR could be related to the observed changes in expression levels of oscillating genes.

BioSeq29

Benchmarking coevolution methods

Rocio Rama Ballesteros[1], Emilie Neveu[1], Nicolas Salamin[1]

[1]*University of Lausanne, Lausanne, Switzerland*

Coevolution is an important component of evolutionary theory and describes the reciprocal changes that occur between biological entities as they depend on each other. It is one of the mechanisms driving biodiversity when interactions occur between entire organisms, but this process can also explain, at the molecular level, how biomolecules interact with each other.

At the molecular level, coevolution can be detected when a modification at one site along the sequence triggers the modification at another site. In this context, coevolution can reveal important information about the function and structure of a protein, because these coordinated changes tend to occur to improve or maintain functional and structural interactions.

Identifying the mechanisms behind the process of coevolution at the molecular scale is still an important question in molecular evolution.
It is known that coevolving sites are spatially coupled, as they tend to be closer in the 3D structure of a protein than independent sites. Nevertheless, correlated evolution between amino-acids in a sequence could also come from other mechanisms that maintain the protein function through evolutionary time scales, such as epistasis or compensatory mutations. This means that coevolving sites could also be located at functionally important sites that are spatially distant.

A variety of methods have been proposed to predict coevolving pairs of sites. Most are only considering the structural constraints, and usually require a multiple sequence alignment (MSA) containing large numbers of sequences. Other methods are also considering the evolutionary constraints that can lead to coevolution and require an MSA and a phylogenetic tree to identify pairs of sites under coevolution.

These two types of methods have never been benchmarked to identify their specific properties and highlight their key differences. In this study, we compared two methods based on these two different approaches: Direct Coupling Analysis (DCA), which is the most widely used approach based on structural constraints, and CoMap, based on evolutionary constraints.
We used simulations to evolve amino-acid sequences along phylogenetic trees by including varying levels of coevolution. We then characterize the ability of each method to detect correctly coevolving sites under varying levels of sequence divergence and different tree topologies. We show that DCA has difficulties to detect coevolving sites if sequence divergence is low, while it is improving with increasing divergence. We show that it is due to the increasing number of the coevolving pattern that are created. In contrast, CoMap finds more false negatives at both low and high levels of sequence divergence, but is accurate at intermediate levels. Finally, we used the two approaches on datasets from the Pfam and Ensemble databases to illustrate the application on real datasets and to understand better the underlying properties of the two methods. Our study shows that the two types of methods give complementary results, which probably highlight the different processes that they are meant to identify.

Compute01

Discriminating Early- and Late-Stage Cancers Using Multiple Kernel Learning on Gene Sets

Arezou Rahimi[1], Mehmet Gönen[1]

[1]*Koc University, Istanbul, Turkey*

Identifying molecular mechanisms that drive cancers from early to late stages is highly important to develop new preventive and therapeutic strategies. Standard machine learning algorithms could be used to discriminate early- and late-stage cancers from each other using their genomic characterizations. Even though these algorithms would get satisfactory predictive performance, their knowledge extraction capability would be quite restricted due to highly correlated nature of genomic data. That is why we need algorithms that can also extract relevant information about these biological mechanisms using our prior knowledge about pathways/gene sets.

In this study, we addressed the problem of separating early- and late-stage cancers from each other using their gene expression profiles. We proposed to use a multiple kernel learning (MKL) formulation that makes use of pathways/gene sets (i) to obtain satisfactory/improved predictive performance and (ii) to identify biological mechanisms that might have an effect in cancer progression. We extensively compared our proposed MKL on gene sets algorithm against two standard machine learning algorithms, namely, random forests and support vector machines, on 20 diseases from the Cancer Genome Atlas cohorts for two different sets of experiments. Our method obtained statistically significantly better or comparable predictive performance on most of the datasets using significantly fewer gene expression features. We also showed that our algorithm was able to extract meaningful and disease-specific information that gives clues about the progression mechanism.

Compute02

An RPCA (Robust Principal Component Analysis) based approach for protein-protein interaction hot-spot prediction.

Divya Sitani[1], Paolo Carloni[1]

[1] *Forschungszentrum Juelich, Germany*

Many disease-associated mutations are located at protein-protein interfaces, causing disrupted or erroneous PPIs (protein-protein interactions). Within these interfaces, only a small subset of residues is crucial for the binding free energy of the protein–protein complex. These residues are known as “hot spots” and they contribute the most to protein-protein binding. Thus, if hot spot residues are mutated, they can severely disrupt PPIs. This makes the identification of such residues extremely important to understand the impact of disease associated mutations on PPIs. Experimentally, hot spots can be found out by using Alanine Scanning Mutagenesis, but it is quite costly and cumbersome. This has led to an increased use of computational methods to predict hot spot residues in the recent years. In our work, we develop a robust principal component analysis based method to predict protein-protein interaction hot spot residues. Our main motivation stems from the fact that we can recover a low rank component A from a highly corrupted training data matrix D using Robust PCA. Once, we recover this noiseless low rank matrix A, we learn a suitable classifier on A and predict whether a residue is a hot spot or not on the test data. To the best of our knowledge, this is the first work that uses a robust principal component analysis based technique for hot spot residue prediction. We use the benchmark HB34 dataset to evaluate the performance of our proposed method. After extensive experimentation, we show that our method is quite efficient and effective for identifying hot spot residues participating in protein protein interactions.

Compute03

Automated extraction of color pattern and anatomical characteristics in dairy cows

Jessica Nye[1], Laura Zingaretti[1], Miguel Pérez-Enciso[1]

[1]*Universitat Autònoma de Barcelona, Spain*

Image analysis has increasingly become an important tool for enhancing productivity in many industries. However, extraction of phenotypes can still be costly and time consuming. We explore automatic image analysis that extracts features from dairy cows which can be used in genetic analysis. In order to remove the unnecessary background information, the current methods require time consuming human inspection. Here, we present and compare a composite method that creates a mask (i.e., removes the background portion of the image) and calculates the proportion of dark and light coloration as well as anatomical features in images of bulls in dynamic backgrounds (e.g., forest, grass, hay, snow, etc.). This composite method combines the supervised algorithm MASK-RCNN, an unsupervised image segmentation approach, and k-means color clustering. The first step identifies the region of interest removing the majority of the background noise, while the second and third steps optimize the identification of the bull and segments the color patterning followed by anatomical measurement. We find a very low discrepancy between the proportion of white and dark between the manual extraction and the composite method (r2 = 0.91); with an immense reduction in data collection time. This automatic composite method greatly improves the efficiency of complex image segmentation and analysis without compromising the quality of the data extracted, making analysis computationally feasible for large data sets.

Compute04

The human heart seen by the eyes of a computer scientist

Marta Garcia-Gasulla[1], Filippo Mantovani[1], Alfonso Santiago[1], Guillaume Houzeaux[1]

[1]*Barcelona Supercomputing Center (BSC-CNS), Spain*

Large computational fluid-dynamics simulations are of paramount importance for advancing biology, medicine, and engineering. Most of these simulations are carried out using large parallel supercomputers and use a massive number of computational resources.

One of the main challenges when developing these simulations is its multidisciplinarity. Their development is always pushing into two different directions: On one hand, trying to model the reality with more precision and details, and on the other one using a larger and larger number of computational resources efficiently. Losing the balance between these two forces implies having a useless simulation or an inefficient code.

To keep this balance, the focus of the high-performance computer scientists needs to be on the programmability and portability of the code. This allows field scientist to focus on functionality when improving science. Looked from another angle, this balance results in being a fruitful and interdisciplinar collaboration among specialized scientists of different fields.

In this talk, I will analyze a biological simulation running on a Tier-0 supercomputer: Marenostrum4, using a state of the art multi-physics parallel code: Alya. With the help of a visual performance analyzer, I will show different factors that cant threat performance in this kind of simulations. While some of the contributions will be technical, I will try to keep the explanation understandable for an educated audience.

The ultimate goal of this talk is to review the challenges of running complex biological simulations on parallel high-performance computing systems and highlight the importance of interdisciplinarity at all levels of research.

Compute05

DLMF: Deep Logistic Matrix Factorization with multiple information integration for drug-target interaction prediction

Sarra Itidal Abbou[1], Hafida Bouziane[1], Abdallah Chouarfia[1]

[1]Université des Sciences et de la Technologie Oran USTO-MB, Algeria

Conducting a new drug development project using wet-lab experiments is quite arduous since it requires an astronomical time and effort as well as a very high cost, from the target identification to a market-ready product. To deal with these issues, a new research domain was born called "Computational Aided Drug Design (CADD)", where various computational (in silico) methods were proposed.

In silico methods for drug-target interactions (DTIs) identification consist in predicting new drug-target association pairs with reasonable cost and accuracy. This crucial step in drug discovery process is performed using various techniques which roughly fall into three major categories: ligand-based, molecular docking-based and chemogenomic-based methods. Since ligand-based methods perform poorly when known ligands for a target is too small and molecular docking cannot be applied when the 3D structure of a target is not available, therefore, the chemogenomic-based methods represent the alternative.

The emerging chemogenomic methods consist in using information from targets and drugs concurrently to infer new drug-target pairs. In order to improve the prediction performance of new pairs, several chemogenomic-based methods released form machine learning techniques, particularly deep leaning were suggested in the literature.

Our proposed method DLMF adopts an hybridization between deep learning and matrix factorization technique in order to boost the prediction performance, since several comparisons have been published showing that matrix factorization yields best results in DTIs prediction compared to the other categories such as network-based methods or bipartite local model methods, along with deep learning which becomes an advanced and significant technique in many research fields.

These state-of-the-art methods achieved good prediction results, however, the majority of them treat only new drug/target prediction cases whereas DLMF represents a new recommendation framework to improve the prediction performance further. In addition, it handles new drug-new target pair prediction settings.

Besides the observed interaction matrix, DLMF integrates multiple information to the prediction model such as, the chemical structure similarity between compounds, the sequence similarity between target proteins and side information for both drugs and targets. The method applies logistic matrix factorization with graph regularization on the observed interaction matrix Y to get two latent feature matrices U and V for both drugs and targets, respectively. Then a drug feature vector u_i, a drug chemical structure similarity vector S_i^d and a drug's side information vector θ_i are concatenated into a unified drug feature vector I_ui and at the same time a target feature vector v_j, a target sequence similarity vector S_j^t, and a target's side information vector θ_j are also concatenated into a unified target feature vector I_vj and then these two feature vectors are fed into a deep learning algorithm for classification. Thereby, when no interaction information is available, such as, the new drug-new target pair case, the side and similarity information is used for classification.

The performance evaluation of our proposed method was performed on the gold standard dataset. using both the area under the curve (AUC) and area under the precision-recall (AUPR) evaluation metrics. The assessment results showed really promising results for DTIs prediction.

Compute15

An atomistic molecular dynamics simulations approach to the study of h-LDHA inhibition

Antonia Vyrkou[1], Simon Allison[1], David Cooke[1], Roger Phillips[1]

[1]University of Huddersfiled, United Kingdom

It has been known that cancer cells reprogram their metabolism in a way that allows them to produce the energy needed for their functions and growth through the conversion of pyruvate to lactate. This reaction is catalysed by the enzyme h-LDHA (human Lactate Dehydrogenase A) which is therefore considered a promising target in cancer pharmacology. Inhibition of this enzyme can deprive cancer cells from their energy supply with insignificant side effects to normal cells.

Ag8 (Ag(NHC)2AgBr2) is an organometallic complex, with increased stability and reduced toxicity, which could be used as an h-LDHA inhibitor. Its inhibitory function is being currently experimentally studied and is the focus of this atomistic molecular dynamics (MD) study.

Molecular dynamics can be proven an invaluable tool in the effort to design and develop novel anticancer pharmaceutical such as h-LDHA inhibitors. Computer simulations, are a relatively fast, inexpensive way to study complex systems under various conditions. Therefore, the results obtained can be an important addition to experimental studies.

In molecular dynamics, the dynamic evolution of the system is produced by allowing the atoms to interact with one another for a period of time under specific conditions. The atomic positions and energies are then obtained by solving Newton’s equations of motion, and the post processing of these data provides information about the structure, properties and behaviour of the system under study. For this study, all molecular dynamics simulations are performed using the open source software GROMACS 5.1.4.

In order to perform MD simulations with meaningful results, structural information and a reliable force field, that is a description of all the inter- and intramolecular interactions, are required. For h-LDHA, this process has been quite straightforward as it has been studied before. Therefore, the initial structure chosen was PDB 1i10, obtained from the protein data bank were h-LDHA is in the form of a tetramer. The force field parameters selected for the description of the bonded and non-bonded interactions were those provided by Amber ff99SB. The initial PDB structured contained oxamate, which for our study was replaced by pyruvate, and the results obtained over a simulation time of 30 ns in the NPT ensemble were in good agreement with the bibliography.

In the case of Ag8, this process is not as easy. First, the initial structure of the molecule had to be described in the form of a PDB file. This was achieved by using the online tool “SMILES Translator and Structure File Generator”. The biggest obstacle encountered, for which the effort to overcome is still ongoing, was to find a reliable set of force field parameters. To that end, three options are considered:

1. The VFFDT software, based on quantum mechanical calculations
2. The acpype tool that creates topologies based on the General Amber Force Field
3. A manual adaptation of the parameters available for compounds of similar structure

POSTERS

BioMed06

File QC Portal: first insight of the sequencing samples deposited at EGA

Aina Jené[1], Dietmar Fernández Orth[1], Claudia Vasallo Vega[1], Babita Singh[1], Jordi Rambla de Argila[1]

[1]Centre For Genomic Regulation, Barcelona, Spain

A picture is worth a thousand words." This adage has been extensively used to define how a complex idea can be conveyed with just a single image, that is how a simple graph can show large amounts of information. The European Genome-phenome Archive (EGA), as a repository of sequencing, variation array and phenotypic data for biomedical research projects, fully supports that proverb. One of the main purposes of EGA is allowing scientists and clinicians to get useful data available for their own analyses at a glance. Thus, EGA is making an effort to facilitate such exploration and is currently implementing some initial quality control (QC) tools for all sequencing data (aligned BAM) and Variant Call Format (VCF) stored files. Raw data (fastq) QC is also analysed by using the FastQC tool. Within the File QC Portal users can visualize several graphs and hence are allowed to check the main characteristics of the file and get an overall idea about its quality and reusability before downloading these. As an example, information regarding the proportion of mapped reads, duplicates, quality distribution can be checked, with an explanation on how to interpret each graph. Moreover, summary statistics and non-sensitive graphs are included, as well. Finally, with the QC Portal we aim to provide universal access to the main characteristics of the data through an interactive and intuitive graphical user interface.

BioMed07

Dietary patterns by food groups are important shapers of the gut microbiota

Angela Sofia Garcia Vega[1], Alejandro Reyes Muñoz[1], Juan Sebastián Escobar Restrepo[2]

[1]Universidad De Los Andes, Bogota, Colombia, [2]Vidarium–Nutrition, Health and Wellness Research Center, Colombia

The Human Microbiome Project and similar efforts around the world opened a promising avenue in personalized nutrition and medicine by showing an intimate association between gut microbiota and human health. One of the most important factors contributing with health is diet, a modifiable aspect of the human lifestyle capable of altering the community dynamics in the gastrointestinal tract. However, associations between intake of nutrients and gut microbiota have been elusive. Recent studies have shed light into these associations by suggesting that diet-microbiota interactions are driven by individual food choices rather than intake of nutrients. In consequence, it is of great interest to analyze diet from the food-choice perspective across different populations. In a culturally diverse population, like Colombia, we found differences in food choices across the country's geography. As such, in this study we explored associations between diet and gut microbiota through both food groups and nutrient intake to test the idea that food choices are more determinant to gut microbiota composition than the analysis of nutrients alone. To test this, we analyzed a community-dwelling cohort of 459 adults from five Colombian cities. The available data for each individual included demographic, clinical, dietary and gut microbiota information, obtained from direct interviews, blood chemistry analyses and 16S rRNA gene sequencing from fecal samples. Using bioinformatic tools and multivariate statistical analyses, we correlated food groups to gut microbiota composition, adjusting the models by other covariates known to affect the gut microbiota, including age, sex and sociodemographic factors. Preliminary results on the gut microbiota, using the Bray-Curtis distance to assess beta diversity, showed that the city of origin is the main driver of the microbial community composition in the studied cohort. Parallel multivariate analysis of dietary data showed that the intake of nutrients did not differ across the surveyed cities. In contrast, food-group consumption significantly varied from each population after independent PCA analysis. Moreover, we found that the abundance of some operational taxonomic units (OTUs) correlated with the participants' city of residence, including bacterial groups previously associated with health and disease, such as Bacteroides and Prevotella. Further analyses are needed to determine if the food-group consumption across cities is capable of explaining the differential bacterial composition of the gut microbiota and their previously determined association to cardiometabolic health.

BioMed08

Multi-omics profiling as an approach to precision oncology in a rare CRC tumor: a case report

Ania Alay[1], David Cordero[1], Lorena Ramírez[1], Ana Vivancos[1], Elena Élez[1], Héctor G. Palmer[1], Xavier Solé[1]

[1]*Bellvitge Biomedical Research Institute (IDIBELL), L'Hospitalet de Llobregat, Spain*

Introduction: Over the last few years, computational biology analysis and integration of large-scale omics data has been increasingly contributing to the field of precision oncology. In this work, we present the study of a patient diagnosed with a very rare metastatic colon choriocarcinoma, who has been extensively profiled using different omics technologies. So far, only 19 cases of colon choriocarcinoma have been reported worldwide according to literature.

Aim: To understand the molecular basis of the patient's disease and identify potentially relevant therapeutic options by applying multiple bioinformatics approaches to the multi-omics data generated from the patient's tumor lesions.

Methods: Two samples from the patient's primary tumor and 5 samples from different metastatic lesions (lymph node, 2 liver, and 2 peritoneum) were obtained from the patient. Additionally, three PDX models were generated from some of these metastases. Comprehensive multi-omics profiling comprising whole-exome sequencing, methylation profiling, copy number arrays, gene expression arrays, and proteomics was performed on all samples. Raw data was preprocessed and analyzed in-house using standard NGS tools (e.g., BWA-mem, GATK, etc.) and R software. Additionally, we used data from large cancer pharmacogenomics assays and catalogs of genetic vulnerabilities to predict potential treatments for the patient under study that were further validated in the in vivo models.

Results: Identification of driver mutations allowed the characterization of the tumor and assessment of better therapeutic options (e.g. BRAF V600E and thus the potential use of BRAF inhibitors). Estimation of tumor's clonal evolution using driver mutations from primary tumor, synchronous metastases, and distant metastases helped identifying drivers responsible of disease progression. Analysis of mutational patterns revealed evidence of a deficiency in homologous recombination, which was further supported by the methylation results (i.e., methylation of BRCA1 and RAD51C promoters) and copy number alterations. Moreover, this deficiency was also validated using experimental approaches, confirming the epigenetic silencing was driving this phenotype. Proteomics data analysis showed a higher activity of the Hippo pathway which was later corroborated through gene dependency analysis. Association of transcriptomic data with pharmacogenomic assays allowed clustering of the samples in a group with better response to EGFR inhibitors, MEK inhibitors, and JAK3 inhibitors. These results were further validated in vivo, corroborating the validity of our approach to identify potential therapeutic strategies using computational biology strategies.

Conclusion: In conclusion, integrative profiling of all the omics layers allowed to identify driver alterations for this disease, generate a phylogenetic reconstruction of the tumor's clonal evolution, as well as to unveil a deficiency in the homologous recombination machinery mediated by epigenetic silencing of key DNA repair regulators. Moreover, the mining of large-scale pharmacogenomics assays also uncovered potential therapeutic strategies that were subsequently validated in vivo. Overall, this study represents a landmark example of how computational biology can contribute to precision medicine and cancer treatment in the near future.

BioMed09

The Mutational Landscape of a Prion-like Domain

Benedetta Bolognesi[1], Andre Faure[2], Mireia Seuma[1], Joem Schmiedel[2], Ben Lehner[2]

[1]Institut De Bioenginyeria De Catalunya, Barcelona, Spain, [2]CRG, Spain

Specific insoluble protein aggregates are the hallmarks of many neurodegenerative diseases. Whether the protein aggregates themselves or other forms of the proteins are toxic to cells is still unclear in many of these diseases. This lack of understanding of the causes of cellular toxicity is reflected in the general failure of multiple therapeutic approaches so far attempted. The causes of this rely mainly on the lack of systematic approaches able to estimate in parallel the effect of mutations on cell viability as well as on protein conformation. Cytoplasmic aggregates of the RNA-binding protein TDP-43 are, for example, observed in 97% of cases of Amyotrophic Lateral Sclerosis (ALS) but their role in the establishment of pathology is unclear. Here, we present our recently developed method that allows us to address this question by systematically mutating TDP-43 and quantifying the effects on cellular toxicity. We generated a dataset of >50,000 mutations in the intrinsically disordered prion-like domain (PRD) of TDP-43 and discovered that changes in hydrophobicity and aggregation potential are highly predictive of changes in toxicity. Surprisingly, however, increased hydrophobicity and cytoplasmic aggregation reduce toxicity. On the other hand, variants that increase toxicity promote the formation of more dynamic liquid-like condensates. The central region of the PRD represents a hotspot of toxicity where genetic interactions in double mutants indicate the presence of specific secondary structures. Therefore, this 'unstructured' region can become partially structured in vivo, promoting aggregation and reducing toxicity. Our results demonstrate that deep mutagenesis is a powerful approach for probing the in vivo structural conformations of disordered regions. Moreover, they reveal that aggregation of TDP-43 is not toxic but actually protects cells, most likely by titrating the protein away from a toxic liquid-like phase. Therefore, promoting rather than alleviating aggregation might actually represent an appropriate therapeutic goal.

BioMed11

Lighting an Evidence Beacon to support disease exploration using the DisGeNET platform

Claudia Vasallo[1], Janet Piñero[2], Sabela de la Torre[1], Frédéric Haziza[1], Babita Singh[1], Dietmar Fernández[1], Laura I. Furlong[2], Jordi Rambla[1]

[1]Centre For Genomic Regulation, Barcelona, Spain, [2]UPF, Spain

A Beacon is a web-accessible service implemented as a communication layer for the federated discovery of genetic variation data and their associated phenotypic effects through querying data stores. More than 100 Beacons from over 40 organizations have been lit since the inception of the Beacon Project as a Global Alliance for Genomics and Health (GA4GH) initiative in 2014. The project keeps growing as international organizations implement Beacons and these are assembled into searchable networks for increased discoverability.

In the Core Beacon implementation, queries are performed against sequence-derived data and include the existence of variants, associated effects, biological or assay-associated metadata, and source. Additionally, biomedical researchers´ needs drive the development of Beacons in many flavours. Evidence Beacon is an emerging flavour of Beacon that aims to implement relevant queries on top of variant knowledge resources such as variant-disease association databases. This will provide context for a certain type of biological annotations, such as disease annotations obtained by text mining the scientific literature.

In this work, we aim at lighting a Beacon for DisGeNET platform. DisGeNET is a knowledge management platform containing information about over 100,000 variants associated with more than 10,000 diseases, traits, and phenotypes. The information in DisGeNET is annotated using community-based standards and a variety of disease vocabularies. Furthermore, the resource maintains an explicit representation of the provenance of the information, which allows the user to trace back to the original source of information. The implementation of this Beacon is based on the specifications for queries that are relevant to researchers interested in variant-disease or genotype-phenotype information and addresses questions such as “What are the diseases, or traits associated with a variant?”, “What are the original database reporting a variant-disease association?”, “What publications(s) support a variant-disease association?”, “What is the confidence of a variant-disease association?”, or “What is the allelic frequency of a variant and its consequence type?”

DisGeNET Evidence Beacon will foster the discoverability of biomedically relevant metadata on disease-related variants and will pave the way for the implementation of Evidence Beacons for similar resources.

BioMed12

In silico prescription of anticancer drugs in single-cell RNA-seq.

Coral Fustero Torre[1], Fátima Al-Shahrour[1]

[1]Spanish National Cancer Research Centre (CNIO), Madrid, Spain

Cancer is the second leading cause of death in developed countries. In recent years, our understanding of cancer has progressed substantially, and the latest advances in diagnosis and treatment have contributed to higher success rates in the battle against many tumour types. However, despite this significant investment and effort, the efficacy of current anticancer treatments is far from satisfactory and many patients still die from the disease.

The idea behind Precision Medicine results from the dependency between the therapy efficacy and the presence of specific genetic alterations. The success of this strategy, relies on its ability to translate accumulating genomic data into actionable treatment options tailored for individual patients. This process requires the identification of a genomic signature from patient tumour samples that will be matched with the most effective therapy.

However, cancer cells are heterogeneous and this heterogeneity can introduce significant challenges in the design of effective treatment strategies. Since tumours can exhibit different sensitivities to cytotoxic drugs among their different clonal populations, targeted treatments could play a role in the appearance of drug resistance mechanisms. Therefore, in order to connect patient-specific data to drug response, clone-specific vulnerabilities must be taken into account.

In recent years, there have been a lot of technological advances in the development of single-cell RNA-seq (scRNA-seq) for the understanding of cell-specific transcriptome profiles. With this technology, researchers are provided with the ability to measure gene expression from individual cells, distinguishing distinct cell subpopulations, and allowing the characterization of cell proportions and composition in disease relevant tissues.

In this work, we present a novel tool for the analysis of scRNA-seq data and in silico prescription of anticancer drugs. For this purpose, we have analysed public single-cell RNA-seq tissue studies coming from the Single Cell Data Portal [1] and generated a subpopulation clustering based on genetic and chemical perturbation signatures by using the available pharmacological profiling data coming from LINCS [2] and Connectivity Map [3]. Treatments have been prioritized depending on their reach or capacity to affect a wider spectrum of the found subclones. Finally, a functional characterization of the distinct pathway mechanisms involved has been applied.

We believe this tool will lead to a better understanding of the biological and therapeutic impact of tumour heterogeneity and could be a valuable resource for personalized medicine.

[1] https://portals.broadinstitute.org/single_cell; [2] Subramanian et al. Cell, 171(6):1437-1452.; [3] https://www.broadinstitute.org/connectivity-map-cmap

BioMed13

Karakitsou

Effrosyni Karakitsou[1], Carles Foguet[1], Pedro de Atauri[1], Jean-Baptiste Cazier, Marta Cascante[1]

[1]Universitat De Barcelona, Barcelona, Spain, [2]University of Bermingham, United Kingdom

The classical approach in cancer research focuses on the identification of specific genomic features, such as single mutations or the role specific genes play in the development and progression of different types of cancer. However, the need to treat cancer as a multifactorial disease and therefore take under consideration the entirety of events and the different levels of regulation that occur is becoming increasingly evident. In addition, accounting for the specific genetic and phenotypic fingerprint of each patient will pave the way towards an accurate stratification of patients, better treatment efficacy and elimination of potential drug side-effects.

Through multi-omic data integration different types of medical and biological datasets can be integrated, so that a more relevant interpretation can be achieved. Genome-Scale Metabolic Models (GSMMs) provide an excellent platform to integrate various omics. They are widely used to computationally simulate the whole of the biochemical pathways and events that take place in a cell at a genome scale and can be used to perform predictive or prognostic studies. Furthermore, statistical modeling and the application of machine learning provides the means to handle the ever-increasing amount of biomedical data and their inherent complexity.

Here we present developments on metabolic modeling tools applied to build an Acute Myeloid Leukemia (AML) specific metabolic model, by integrating transcriptomic and metabolomic data from different AML cell-lines and patients. The flux distributions provided by the simulations together with the meta-data of the patients were used to build classifiers for AML that could be used in clinical decision making. The proposed pipeline is a first step towards the use of GSMMs and machine learning in the delivery of personalized cancer therapy and patient stratification.

Acknowledgements: European Commission (HaemMetabolome, EC-675790), Catalonia Government (2017SGR1033), and Spanish Government- MICINN (SAF2017-89673-R)

BioMed14

Prediction of the Effects of Single Amino Acid Variations on Protein Functionality with Annotation Centric Modeling

Fatma Cankara[1], Tunca Doğan[2]

[1]Middle East Technical University, Ankara, Turkey, [2]EMBL-EBI / Hacettepe University, Turkey

Whole-genome and exome sequencing studies have indicated that genomic variations may cause deleterious effects on protein functionality via various mechanisms. It has been reported that single nucleotide variations that alter the protein sequence (and thus, the structure and the function), namely nsSNPs, are highly associated with genetic diseases in human. These variations also affect the receptor-ligand interactions, which is one of the key reasons that many of the early drug candidates fail in clinical trials. In this study, we propose a new methodology to collect and organize the information related to the effects of sequence variations from various biological databases and to utilize this information in a machine-learning based system to predict the disease-causing capabilities of mutations with unknown consequences.

The studies aiming to predict the effect of sequence alterations in proteins often exploit sequence conservation (mostly using alignment) and 3-D structural information. In this study, we took a different perspective and evaluate these variations' disease-causing capabilities using an annotation-centric focus (i.e., domains, motifs, and other sequence features), with the aim of complementing conventional approaches in the literature. Functional regions/sites of proteins are the parts, where a sequence variation may have a more significant consequence and should be evaluated accordingly. In the proposed methodology, we make use of a variety of descriptive features including:

i.

a) The correspondence between the mutated site and different protein sequence feature annotations (obtained from the UniProt database) such as taking part in disulfide bonding, nucleotide binding, zinc fingers, glycosylation, helix, repeats, etc. (30-dimensions, binary).

b) For the cases where the mutation does not perfectly correspond to the annotated sites on the sequence, the atomic distances (on the 3-D structure) between the mutated residue and the annotated residue (based on alpha-carbons) are incorporated in an additional 30-D vector with real values (angstrom).

c) The information about InterPro/Pfam domain entry, where the mutation resides in (1-D, categorical).

ii. Mutated residue's structural information that we deduce from PDB structures by aligning the protein sequence with the corresponding PDB structure's sequence and by categorizing the mutated residues based on their accessible surface area as core, surface or interface (1-D, categorical).

iii. Physicochemical properties of the mutation obtained from the widely accepted Grantham Matrix's 3 distance scores: the change in polarity, composition and molecular volume, upon the occurrence of the variation (3-D, real-values).

We employed decision tree (DT) and random forest (RF) classifiers, first, to train our models (using the mutations with known consequences from UniProt, Clinvar and PMD, a total of 108072 mutations), and then, to predict the effect of unknown variations by categorizing them either as disease-causing or neutral, by querying the finalized 65-D feature vector of each variation, on our prediction models. Currently, we are in the process of running our models on widely accepted benchmark datasets from previous studies to calculate our performance, and to compare them with the state-of-the-art. Finally, we plan to combine our method with the state-of-the-art methods, which employ different featurization approaches, to maximize the prediction performance in an ensemble-based tool.

BioMed15

DIFFERENTIAL EXPRESSION ANALYSIS OF LYSOSOMAL STORAGE RELATED GENES IN GLIOMAS

Gerda Cristal Villalba Silva[1], Ursula Matte[1]

[1]UFRGS, Porto Alegre, Brazil

Introduction: Defects in activity and transport of lysosomal hydrolases lead to the accumulation of lysosomal metabolites, causing lysosomal storage disease (LSD). The involvement of lysosomes also has been described in tumors, related to cell proliferation and signaling processes, microbial killing, cytotoxic killing, induction of angiogenesis, cell adhesion and metastatic processes. Aims: This work aims to investigate the differential expression of lysosomal genes and autophagy-related genes in brain tumors. Data from patients with different brain tumors were collected and analyzed through the R2: Genomics Analysis and Visualization Platform. Gene expression analysis was performed on 50 gliomas with five pathological diagnosis. Sample tracks were divided in recurrence of progression and description of the tumor. All differentially expressed genes were correlated with survival rates using Kaplan Meier. Results: From 113 selected genes (KEGG pathway: lysosome), 24 were up-regulated and 6 were down-regulated. Heatmap analysis showed that 30 genes have a distinct pattern for grade IV glioblastoma multiforme. Survival analysis showed that cathepsin expression correlates with poor survival, as previously shown for other tumors. In addition, similar results were found for different lysosomal hydrolases, such as Beta-Manosidase, Galactosidase, Glucorunidase beta, and, Hexosaminidase, and autophagy-related genes. Conclusion: Studying the lysosomal signature in tumors may help understand the mechanisms by which neurological dysfunction occurs both in patients with neurological tumors and in patients with lysosomal storage diseases. Besides, differentially expressed genes may serve as prognostic biomarkers and therapeutic targets. Also, may improve diagnosis and lead patients to a more adequate treatment.

BioMed16

Genomic segments with different DNA repair dynamics

Hanna Kranas[1], Joan Frigola[1], Ferran Muiños[1], Abel Gonzalez-Perez[1], Nuria Lopez-Bigas[1]

[1]Institute For Research In Biomedicine (IRB Barcelona), Barcelona, Spain

Mutation rates are known to be dependent on different genomic features, such as transcription, nucleosome occupancy, the presence of bound transcription factors, histone marks, and others. We know that mutations are an effect of damage acquisition and repair activity, both of which can be distributed differentially, and therefore we aim to find out which - the distribution of the damage or the activity of repair - is responsible for the influence of each feature on the distribution of mutations. To explore this issue, we leverage publicly available data from experiments mapping UV light damage at different time points after UV exposure, and use it to infer the DNA repair activity between them.

Here, we present a novel computational method based on an adapted version of chromHMM that segments the genome into DNA repair states. By pairing it with Gradient Boosting regressors and classifiers, we study how different repair dynamics characterized by different chromatin features shape the damage landscape.

As a result, we are able to obtain genomic features with the highest impact on damage formation, efficiency of repair or mutation deposition and we provide a framework for studying the dynamics of the DNA repair machinery at various parts of the genome. Additionally, this pipeline will allow us to explore if the damage caused by other mutagens and its repair are characterized by the same or different repair dynamics, and the predicted repair states projected to other tissues potentially can be used to predict the trajectory of damage and acquisition of mutations over time.

BioMed17

ResMarkerDB: a database of biomarkers of response to antibody therapy in breast and colorectal cancer

Judith Pérez Granado[1], Janet Piñero[1], Laura I. Furlong[1]

[1]*FIMIM - Fundació Institut Mar D'investigacions Mèdiques, Barcelona, Spain*

The clinical efficacy of therapeutic monoclonal antibodies for breast and colorectal cancer has greatly contributed to the improvement of patients' outcomes by individualizing their treatments [1]. However, primary or acquired resistance to treatment reduce its efficacy [2,3]. In this context, the identification of biomarkers predictive of drug response would support research and development of alternative treatments [4]. Currently, several molecular biomarkers of treatment response for breast and colorectal cancer have been described. However, this information is scattered across several resources and not properly integrated hindering its potential use. Therefore, there is a need for resources that offer biomarker data in a harmonized manner to support the identification of actionable biomarkers of response to treatment in cancer. ResMarkerDB was developed as a comprehensive resource of biomarkers of drug response in colorectal and breast cancer. It integrates these data from existing repositories, and new data extracted and curated from the literature (referred as ResCur). The database contains more than 500 biomarker-drug-tumour associations. ResMarkerDB provides a web interface (http://www.resmarkerdb.org) to facilitate the exploration of the current knowledge of biomarkers of response in breast and colorectal cancer. It aims to enhance translational research efforts in identifying actionable biomarkers of drug response in cancer.

References

[1] Chiavenna,S.M. et al. (2017) State of the art in anti-cancer mAbs. J. Biomed. Sci., 24, 15.
[2] Pruneri,G. et al. (2016) Biomarkers for the identification of recurrence in human epidermal growth factor receptor 2-positive breast cancer patients. Curr. Opin. Oncol., 28, 476–483.
[3] Bronte,G. et al. (2015) New findings on primary and acquired resistance to anti-EGFR therapy in metastatic colorectal cancer: do all roads lead to RAS? Oncotarget, 6.
[4] Bossuyt,P.M. and Parvin,T. (2015) Evaluating Biomarkers for Guiding Treatment Decisions. EJIFCC, 26, 63–70.

Funding

This work was supported by "Instituto de Salud Carlos III- Fondo Europeo de Desarrollo Regional" [grant numbers: PIE15/00008, CP10/00524, CPII16/00026];, "Instituto de Salud Carlos III- Fondo Social Europeo" [FI18/00034]; and the European Commission Horizon 2020 Programme 2014- 2020 under grant agreements MedBioinformatics [grant number: 634143] and Elixir-Excelerate [grant number: 676559]. The Research Programme on Biomedical Informatics is a member of the Spanish National Bioinformatics Institute, "Plataforma de Recursos Biomoleculares y Bioinformáticos-Instituto de Salud Carlos III" [grant number: PT13/0001/0023], of the PE I+D+i 2013-2016, funded by "Instituto de Salud Carlos III" and "Fondo Europeo de Desarrollo Regional". The "Departamento de Ciencias Experimentales y de la Salud" is a "Unidad de Excelencia María de Maeztu", funded by the "Ministerio de Economía y Competitividad" (reference number: MDM2014-0370)

BioMed18

Exploring the genetic architecture of Major Depression: low agreement between the results of candidate gene studies and GWAS

Judith Pérez Granado[1], Janet Piñero[1], Laura I. Furlong[1]

[1]FIMIM - Fundació Institut Mar D'investigacions Mèdiques, Barcelona, Spain

Major depression (MD) is a prevalent mental disorder and the leading cause of impairment worldwide, being more common in women and with a significant health and socioeconomic impact. MD arises from a complex interplay between multiple genetic and environmental factors. MD is diagnosed by the presence of a set of symptoms, and no biomarker or laboratory test is available to aid diagnosis and treatment selection. While the main treatments for MD are both, psychotherapy and antidepressant medication although pharmacological treatment is currently not effective in 40% of patients. Moreover, for those patients that respond to treatment, there is an important delay in the onset of therapeutic effect.

The genetic susceptibility of MD is 31-42%, estimated by twin studies. In the last years, thanks to the availability of large cohorts and the use of minimal phenotyping, Genome Wide Association studies (GWAs) have enabled the identification of risk loci for MD. In addition to GWAs, genetic architecture of MD has been studied by candidate gene studies and targeted sequencing. It is expected that a better understanding of the genetic contribution to MD may allow a better understanding of the molecular mechanisms of the disease, and therefore lead to the identification of potential drug targets. To support these goals, we aim to develop a knowledge platform that gathers and organizes the genomic alterations associated to MD and allows a variety of analysis to assess the functional consequences of these genomic alterations.

Hereafter we present the results of an initial characterization of the landscape of MD genomic alterations, including the assessment of the pattern of tissue expression of MD genes, analysis of functional consequences of sequence variants and pathogenic prediction, and network analysis of risk variants.

A careful selection of the terminology was required to identify MD studies, using controlled vocabularies and term expansion by semantic approaches. On one hand, we collected MD risk loci from the public repositories GWAS catalog and GWAS db. On the other hand, DisGeNET was used to retrieve genes and variants associated to MD from more than 400 publications mostly reporting candidate gene studies. Overall, more than 400 genes and 500 genomic alterations were obtained as associated to MD. The majority of the genomic alterations are located in non-coding regions. Seventy percent of genomic variants located in coding regions are mapped to genes significantly expressed in brain regions. Strikingly, the overlap of genes and variants identified by GWAs and candidate gene studies is extremely low. We analyze and discuss the possible reasons of the discrepancies among these different experimental approaches in MD.

Immunogenomic characterization of renal cell carcinomas

Laia Bassaganyas[1], Philip S Smith[1], Bryndis Yngvadottir[1], Eguzkine Ochoa[1], Eamonn M Maher[1]

[1]*University Of Cambridge, Cambridge, United Kingdom*

Renal cell carcinoma (RCC) comprises a heterogeneous spectrum of tumours typically presenting a T-cell-inflamed (hot) phenotype. RCC is a highly therapy-resistant cancer, but immune checkpoint inhibitors (ICIs) have shown improved effectiveness in a subset of patients. However, the success rate is still limited to only a small fraction of cases, and predicting them is not yet feasible. In contrast to other hot tumours, RCC is characterized by harbouring a low burden of somatic mutations, indicating that other genomic players might be predicting immunotherapy response. Recent studies show that aneuploidy may promote tumour immune evasion, which suggests its important role in shaping the immunogenic microenvironment, but the effect of other structural variants (SVs) remains unexplored. Here, in order to decipher the interplay between the cancer genome and the immune system in the most common RCC subtypes (clear cell RCC –ccRCC-, papillary RCC –pRCC-, and Chromophobe –ChRCC), we conducted an integrative analysis including WGS/WES and whole gene-expression profiling using samples from the PCAWGS, TCGA and additional RCC cohorts. Aneuploidy levels were quantified for each sample by computing a Broad-CNA Score and a Focal-CNA Score based on the number, amplitude and length of copy-number alterations (CNAs). These scores were correlated with the tumour mutational burden (TMB), the number of SVs and transcriptomic results interpreting the tumour immune composition. Results clearly show distinct immune profiles for the three RCC subtypes, with ccRCC presenting significant higher levels of multiple immune cell types infiltration. Correlation analyses showed that genomic alterations interfere differently in tumour immunity depending on the histological type, and that the direction of association is not always consistent with the pan-cancer perspective. In addition, unsupervised clustering indicates the presence of specific immunogenomic subgroups within each tumour subtype. Overall these data highlights the complex relationship between the genome and the immune system, strengthen the importance of tumour origin within the kidney malignancies, and provides a unique portrait of the tumour and its microenvironment that should be considered to stratify patients and to promote more effective immunotherapy approaches.

Targeting Intratumoral Heterogeneity with in silico Drug Prescription Tools

María José Jiménez-Santos[1], Elena Piñeiro-Yáñez[1], Gonzalo Gómez-López[1], Fátima Al-Shahrour[1]

[1]Spanish National Cancer R3esearch Centre (CNIO), Madrid, Spain

Precision medicine is an emergent field whose aim is to develop improved prognostic, diagnostic and therapeutic strategies for each patient according to thousands of people's clinical and genomics data [1]. Precision medicine is particularly interesting in oncology, since cancers are characterized by a high genomic heterogeneity among different tumors (intertumoral heterogeneity) and within the same cancer (intratumoral heterogeneity, ITH) [2]. Recently, some in silico drug prescription tools have emerged to prioritize tumor's specific genomic alterations with matched therapies and candidate drugs [3]. Nevertheless, these resources often use a therapy administration strategy based on bulk results and ignore ITH.

ITH refers to the existence of different cell populations within the same tumor and can be spatial (variability in different locations) or temporal (variability over the temporal evolution of the tumor) [4]. ITH has been revealed as a key factor in cancer patients' outcome contributing to higher lethality, therapy failure and drug resistance [5]. Thus, in silico drug prescription performance could be improved by integrating ITH information into preexistent tools.

In this work, we used PanDrugs (www.pandrugs.org), an in silico drug prescription tool developed in our laboratory, to design anticancer treatment regimens considering temporal and spatial ITH. PanDrugs is a bioinformatics platform that identifies druggable genomic alterations and prioritizes drug therapies based on clinical, biological and pharmacological evidence [6]. This study applied PanDrugs to mutational profiles obtained by temporal and spatial ITH dissection in acute myeloid leukaemia (AML) and non-small cell lung cancer (NSCLC) patients respectively.

To assess the therapeutic impact of temporal ITH, we used the whole genome sequencing data of the primary and relapse tumor of an AML patient reported by Ding et al. [7]. We were able to propose approved drugs to hit minority clones detected at early stages that were positively selected after chemotherapy. These therapies could have been administered after the standard treatment or in combination to delay or prevent relapse. Moreover, we analyzed two NSCLC patients sequenced by Jamal-Hanjani et al. [8] in order to identify therapeutic differences when spatial ITH is considered. PanDrugs' results suggested to simultaneously target clonal (trunk) mutations with a drug and subclone populations with additional specific therapies. Finally, we verified that taking into account the ITH can increase the number of druggable genes detected and expand the antitumoral drug arsenal with respect to a bulk analysis.

In summary, these results indicate that it is possible to identify drugs or combinations capable of covering the clonal diversity of the tumor. This strategy could help to target minority clones that would otherwise be favored by the administration of the treatment and cause relapse. Moreover, ITH dissection could be a very valuable strategy to increase the therapeutic options in cancer patients.

References

1.Biankin. Nat Genet. 2017;49(3):320-321.
2.International Cancer Genome Consortium et al. Nature. 2010;464(7291):993-8.
3.Tamborero et al. Genome Med. 2018;10(1):25.
4.McGranahan and Swanton. Cell. 2017;168(4):613-628.
5.Saeed et al. Int J Cancer. 2019;144(6):1356-1366.
6.Piñeiro-Yáñez et al. Genome Med. 2018;10(1):41.
7.Ding et al. Nature. 2012;481(7382):506-10.
8.Jamal-Hanjani et al. N Engl J Med. 2017;376(22):2109-2121.

BioMed37

Functional variant prioritization in rare diseases using a mechanistic approach

Ana M. Pérez-Gutiérre[1], **Rosario Maria Carmona Muñoz[1]**, Virginia Aquino-Quintans[1], Javier Perez Florido[2], Kinza Rian[1], Maria Peña-Chilet[3], Joaquin Dopazo[3]

[1]Clinical Bioinformatics Area (FPS), Spain, [2]Servicio Andaluz de Salud (SAS), Spain, [3]Clinical Bioinformatics Area (FPS); CIBERER, Spain

One of the main problems with Rare Diseases (RD) is the still large number of undiagnosed cases, in which the expected simplicity is not found. Although approximately 80% of RDs have a genetic cause, the wide range of phenotypic variability, besides the lack of knowledge of the responsible mechanisms of diseases, makes it difficult to find new disease-causing genes and to determine new molecular therapeutic targets. Indeed, prioritization strategies used so far, targeting one gene at a time, failed to detect the causative genes in many cases, resulting in a list of Variants of Unknown Significance (VUS). Conversely, a more systems biology oriented approach could render better results. Therefore, the detection of systematically affected pathways using a mechanistic mathematical model could provide a new approach for candidate gene prioritization that allow the identification of new disease genes. In addition, they can be used to predict the potential consequences that perturbations (mutations or changes in the expression) of the proteins that compose the pathway can have in healthy tissue over the individual circuits that trigger cell functions.

As a case study, we used genomic data from a cohort of 130 individuals presenting some type of rare disease. We prioritized variants according to ACMG criteria and assigned these variants a level of evidence. In order to evaluate the effect of those variants in the human healthy tissue we performed a Mechanistic Pathway Activity (MPA) analysis using Hipathia algorithm. Using this tool, signaling circuits are defined using KEGG pathways. Gene expression data obtained from healthy tissues transcriptome data from GTEx, were used as proxies of the protein activities of the reference data.

In order to evaluate the potential effect that mutations harbored by each individual would have over the signaling circuits we performed in silico knockout obtaining a score of relevance of each circuit node. This score will help to prioritize the impact that Loss of Function (LoF) mutations may have into a given pathway in a given tissue.

In this work we presented a new methodology to prioritize variants that uses a mechanistic mathematical model of signaling pathway activity, as a proxy for cell functional activity, in contrast to other methods that only consider genes individually. Using this approach, we can obtain the specific signaling circuits associated to the corresponding HPO terms and gain insights into rare diseases mechanisms. Moreover, this algorithm can be implemented in variant prioritization tools helping us to identify new candidate genes and prioritize VUS.

BioSeq07

A large-scale RNA-seq screen reveals novel transcriptional regulators in Neurospora crassa

Aileen R. Ferraro[1], Zachary Lewis[1]

[1]*University Of Georgia, Athens, United States*

DNA-based processes in eukaryotes are controlled by histone modifications that vary in type and location. These modifications regulate transcriptional activity by establishing different types of chromatin domains: transcriptionally active and transcriptionally silent. Regulation of establishment and maintenance of transcriptionally silent domains is poorly understood in in any organism. The model filamentous fungus Neurospora crassa provides an attractive system to study these processes due to its conserved epigenetic machinery, genetic tractability, and the presence of a whole-genome knockout collection. Here we present a targeted RNA-seq screen that reveals novel transcriptional regulators.

BioSeq08

Impact of the sequencing coverage in metagenetic and metagenomic studies

Amandine Bertrand[1], Emilie Detry[1], Cécile Nouet[1], Marc Hanikenne[1], Denis Baurain[1]

[1]Lille University / Liège University, Liège, Belgium

Over the past years, the number of microbiome studies using metagenetic and metagenomic sequencing methods has increased substantially. Those methods are indeed useful to get information about the diversity of organisms (and of genes) present in a large range of sample types, going from medicine to ecology. However, the impact of some technical factors on the results of those methods (generally taken for granted) remains poorly known, such as the coverage used to sequence the metagenetic amplicons. How to know which sequencing coverage (i.e., sequencing depth) is the best suited for a given sample and/or biological question? What are the consequences of a too low or too high sequencing coverage on subsequent interpretations?

We investigated these issues in the wider context of a study of the microbiome of the rhizosphere of Arabidopsis halleri. This plant is hypertolerant and hyperaccumulator of large quantities of metals, more specifically of zinc and cadmium. A. halleri is able to live on both polluted and unpolluted soils. Here, we wanted to identify the whole diversity of microorganisms living near the root of this plant and which could interact with it, especially in relation with metal hypertolerance and hyperaccumulation. To this end, we sampled on a zinc-polluted site in Belgium (Prayon, in the region of Liège), following a sampling scheme allowing us to make a number of useful contrasts.

Different sequencing coverages were used to sequence three types of amplicons of the small subunit (SSU) rRNA gene (16S), allowing us to identify organisms of diverse taxonomy, each from three subsamples of a single sample point. Those coverages were obtained both by increasing the sequencing depth through additional runs and through in silico pooling and by in silico reduction. By this approach, we wanted to have an idea of the minimum coverage required to get an accurate picture of the diversity in all our samples. Different approaches using both statistics and phylogeny inference were used to address this question.

Hence, we observed that the choice of the coverage is not straightforward. When analyzed separately using discovery curves, any specific coverage seems adequate. Yet, more and more operational taxonomic units (OTUs) are discovered with each deeper coverage. Besides, as different primer pairs amplify different organisms, this factor also influences the number of OTUs obtained for one given coverage. Finally, we showed through phylogeny that additional phyla were discovered when increasing the sequencing depth and that new OTUs were thus not only chimera or closely related strains of already observed OTUs. Keeping in mind those observations, one should consider the impact of both sequencing coverage and primers on the inferred diversity, especially in the context of comparative studies.

BioSeq09

The Functional Iso-Transcriptomics analysis framework to assess the functional impact of alternative isoform usage

Lorena de la Fuente[2], Manuel Tardáguila[1], Hector del Risco[1], Angeles Arzalluz[2], Francisco Pardo Palacios[2], Pedro Salguero[2], Cristina Martín[2], Sonia Tarazona[2], **Ana Conesa[1]**

[1]University of Florida, Gainesville, United States, [2]Centro de Investigaciones Principe Felipe, Spain

Post-transcriptional mechanisms such as Alternative Splicing (AS) and Alternative PolyAdenylation (APA) regulate the maturation of pre-mRNAs and may result in different transcripts arising from the same gene, increase of diversity and regulation capacity of transcriptomes and proteomes. AS and APA has been extensively characterized at the mechanistic levels but to a lesser extent in terms of functional impact. While functional profiling is widely used to characterize the functional relevance of gene expression at genome-wide level, similar tools at isoform resolution are missing. Short-reads have limitations to reconstruct full-length transcripts and hence accurately study alternative isoform expression. As single molecular sequencing technologies become more accurate and allow for direct sequencing of full-length transcripts, novel tools are needed leverage the information potential of these platforms to study the functional consequences of alternative transcript processing. Here we present a novel computational framework for Functional Iso-Transcriptomics analysis (FIT), specially designed to study isoform (differential) expression from a functional perspective. This framework consists of three bioinformatics developments. SQANTI is used to define and curate expressed transcriptomes obtained with long read technologies. SQANTI categorizes full-length reads, evaluate their potential biases and removes low quality instances to rendering high-confidence full-length transcriptomes. The IsoAnnot pipeline combines multiple databases and function prediction algorithms to return a rich isoform-level annotation file of functional domains, motifs and sites, both coding and non-coding. Finally, the tappAS software introduces novel analysis methods to interrogate different aspects of the functional relevance of isoform complexity. The Functional Diversity Analysis (FDA) evaluates the variability at the inclusion/exclusion of functional domains across annotated transcripts of the same gene. The Differential Analysis Module evaluates the relative contribution of transcriptional (i.e. gene level) and post-transcriptional (i.e. transcript/protein levels) regulation on the biology of the system. Measures of isoform relevance such as Minor Isoform Filtering, Isoform Switching Events and Total Isoform Usage Change contribute to restricting analysis to biologically meaningful changes. Finally, novel methods for Differential Feature Inclusion, Co-Feature Inclusion, and the combination of UTR-lengthening with Alternative Polyadenylation analyses carefully dissects the contextual regulation of functional elements resulting from differential isoforms usage. tappAS complements statistical analyses with powerful browsing tools and highly informative gene/transcript/CDS graphs.

We applied the FIT framework to the analysis of differentiating mouse neural cells, whose transcriptome was defined by PacBio and Illumina. Using FDA we found that 90% of multi-isoform genes presented variation in functional features at the transcript or protein level. The Differential Analysis module revealed a high interplay between transcriptional and post-transcriptional regulation, where alternative transcript processing acts the main driver of mechanisms such as vesicle trafficking, signal transduction and RNA processing. The Differential Feature Inclusion analysis showed that alternative transcript processing increased the availability of functional features in differentiated neural cells, and is a mechanism for altering gene function by changing cellular localization and binding properties of proteins (NLS, transmembrane domains and DNA binding motifs). A number of these findings were experimentally validated.

We anticipate the FIT framework will help to advance our understanding of the functional significance of alternative transcript processing.”

BioSeq10

Measuring isoform co-expression in single-cell RNAseq successfully decodes splicing coordination as a key determinant of neural cell-type identity

Ángeles Arzalluz-luque[1], Sonia Tarazona[1], Ana Conesa[2]

[1]Universidad Politécnica De Valencia, Valencia, Spain, [2]University of Florida, United States

Single-cell RNAseq studies have mainly focused on the discovery and characterization of new cell types, leaving areas such as isoform expression dynamics unexplored. Furthermore, splicing has almost solely been analyzed in an exon-wise manner. Instead, we hypothesize that splicing can be understood as a regulatory layer that acts coordinately to change the biological properties of entire transcripts and proteins, depending on the context of the cell. This mechanism can be observed as a series of transcript modules that present co-variation across cell types and states, in a manner that is independent of gene-level expression. In this work, we have developed the first computational pipeline that makes use of single-cell RNAseq data to infer such relationships, while also overcoming three main limitations of previous analyses: (1) assessment of differential expression across multiple groups, avoiding pairwise comparisons, (2) new correlation metric that reduces the negative effect of single-cell noise and captures relevant co-expression relationships, and (3) new clustering strategy to group transcripts with similar expression profiles, inferring differential isoform usage and co-usage beyond pairwise analyses.

Our method achieves this by introducing three novel aspects to single-cell bioinformatics: (1) unlocking ANOVA for single-cell data via the introduction of ZINB-WaVE observation weights for multi-group DE inference, (2) computing correlations on the distribution of expression values of an isoform in a cell type, i.e. using the 10 expression deciles for each cell type instead of all expression values, and (3) hierarchical clustering of transcripts followed by a novel, semi-supervised strategy for correlation-based cluster curation that maximizes expression profile similarity among cluster members. As a result, we are able to obtain genes with multimodal differential isoform usage (i.e. differential splicing) by selecting the sets of isoforms that belong to the same gene, but are assigned to different clusters and therefore are expressed in a different modality across cell types. This allows us to extract co-differentially spliced sets of genes, namely the groups of differentially spliced genes whose isoforms are assigned to the same clusters.

In order to test the biological implications of this phenomenon for neural cell type identity, we performed an isoform-level analysis of 24 genes sharing isoforms between a cluster including transcripts with high expression in neural endothelium, and a second cluster including transcripts with high expression in both GABA-ergic and glutamatergic neurons. We found that isoforms from these genes expressed in neurons had consistently longer 3'UTRs, and moreover, that these UTRs were highly populated with miRNA and RBP binding motifs, providing evidence of implications for transcript expression regulation and splicing. In addition, we observe that endothelial isoforms tend to suffer from skipping of exons that are included in neurons, often leading to domain and NLS signal loss. Together, these results point towards a coordinated action of splicing that provides the neural isoforms of these genes with higher functional complexity, both at the protein and transcript levels. Interestingly, our results show that these genes are frequently associated to cytoskeleton remodelling and vesicle transport, opening an interesting path for further analysis and biological interpretation.

BioSeq11

Single Cell Expression Atlas: Systematic Analysis and Visualisation of Single Cell RNA-Seq

Anja Füllgrabe[1], Irene Papatheodorou[1]

[1]*EMBL-EBI, Cambridge, United Kingdom*

Recent advances in technologies for RNA sequencing at single cell resolution have led to a large, increasing number of data sets available within the public domain. Single cell RNA-Seq data sets currently available within archives, such as ArrayExpress and Gene Expression Omnibus (GEO), evaluate a wide variety of different experimental hypotheses, in different species, experimental conditions or diseases. There is an important need to systematically analyse and organize this volume of data, enable informative searches and integration across available experiments that can lead to a better understanding of similarities and differences between cell types, tissues and species. Crucially, effective integrative analysis of scRNA-Seq data sets will support the development and assembly of emerging atlassing projects, such as the Human Cell Atlas and the Fly Cell Atlas, amongst others.

EMBL-EBI's Single Cell Expression Atlas (http://www.ebi.ac.uk/gxa/sc) is an added value database and web-service that annotates, re-analyses and displays gene expression data from single cell RNA-Seq studies. It is part of EMBL-EBI's resource for gene and protein expression, the Expression Atlas (http://www.ebi.ac.uk/gxa/). The web-service can be searched by genes within or across species to reveal experiments, tissues and cell types where this gene is expressed or under which conditions it is a marker gene. Within each study, cells can be visualised using a pre-calculated t-SNE plot and can be coloured by different metadata or by cell clusters, based on gene expression. Gene expression across different cells is also displayed using a t-SNE-based visualisation. Within each experiment, there are links to downloadable files, such as RNA quantification matrices, clustering results, reports on protocols and associated meta-data, such as assigned cell types.

All data sets are re-analysed in-house, using scalable and cloud-enabled workflows. These comprise tools from widely used pipelines such as Scanpy and Seurat that have been decomposed in their corresponding steps and can be used in an interchange-able manner across different workflow systems. We offer a downloadable, cloud-enabled interactive environment in Galaxy that distributes the full analysis workflow performed on data sets in Single Cell Expression Atlas, while facilitating the construction of further workflows to enable benchmarking.

Data curation involves a semi-automatic process of identifying the experimental factors, such as cell types, diseases or perturbations, annotating metadata with Experimental Factor Ontology terms (EFO) and describing the experimental comparisons for further processing. Cell types are annotated with terms from the Cell Type Ontology. Within the full resource, Expression Atlas, we provide results on over 3,500 experiments that include over 150,000 assays from 60 different organisms. The data sets cover over 100 cell types from the Cell Ontology and over 700 diseases represented in the EFO. The single cell transcriptomics part now contains over 500,000 cells, across 98 studies and 11 species, such as Homo sapiens, Mus musculus, Drosophila melanogaster. Single Cell Expression Atlas includes the first data sets available from the Human Cell Atlas.

BioSeq12

A new functional gene annotation system based on STRING protein interaction network clusters for gene set enrichment analysis

<u>Annika Gable</u>[1], Damian Szklarczyk[1], David Lyon[1], João Rodrigues[1], Christian Von Mering[1]

[1]*University Of Zurich, Switzerland*

Background
Functional gene set enrichment using gene annotations such as Gene Ontology terms has become a crucial method to interpret the results of genomics and quantitative proteomics experiments. While a plethora of different functional gene set enrichment methods are available today, all of these rely on curated and annotated gene sets, or in a few cases on gene sets derived from previous experiments or predictions.
Here, we present a functional gene set enrichment of the functional class scoring type that uses gene sets derived directly from the protein-protein interaction database STRING.

Description
For each of the 5090 organisms represented in STRING, we clustered the protein-protein interaction network into hierarchical clusters, retaining the clusters at all levels in the hierarchy. We named these hierarchical clusters by combining the closest matching annotations from functional and domain annotation databases. The generated clusters can now be used as a protein-network-based functional annotation resource covering all 5090 organisms without the need for manual curation.

We found that these hierarchical protein clusters correspond very well to the known functional annotations. The naming process and visualization created an intuitive way to browse the hierarchy, and in which poorly annotated hierarchical clusters can be identified in order to discover new functional clusters or pathways.

We then implemented a randomization-based functional enrichment method for interpreting transcriptomics or proteomics experiments on the STRING website. This method works directly on log fold changes or log p-values and thus does not require setting a significance threshold. The experimental values can be tested for enrichment for both widely-used functional annotations such as Gene Ontology as well as on our new hierarchical clusters.

Conclusions
We have hierarchically clustered the protein-protein interactions for all organisms in STRING. These clusters complement known annotations and allow us to identify potential new pathways in an automated fashion. This could provide new insights into protein functions, especially in less studied taxa.

The new hierarchical clusters can now be enriched for alongside the established gene annotations of Gene Ontology and others. This service is provided within the new functional enrichment method now available as part of STRING version 11 at string-db.org.

BioSeq13

Future of Sharing Genomics Data: Glimpse of EGA's initiatives

Babita Singh[1], Claudia Vasallo[1], Dietmar Fernández Orth[1], Jordi Rambla de Argila[1]

[1]*Centre For Genomic Regulation, Barcelona, Spain*

The European Genome-phenome Archive (EGA) is a repository that facilitates access and management for long-term archival of human biomolecular data. As the omics community awareness of data sharing and reproducibility increases, complex services and granular solutions are needed from repositories such as EGA. While EGA is committed to incorporate state-of-the-art guidelines issued by international committees working in genomics such as Global Alliance for Genomics and Health (GA4GH), it also holds the responsibility to create awareness about such data practices to other researchers and institutions. With exponential growth in genomic data generation, we must follow common practices/guidelines to maximize data generation and distribution to researchers without jeopardizing the security or release of personal information. This responsibility requires the implementation of fair and up-to-date practices in the field of genomics data sharing that integrate an ethical, legal, and social perspective. We aim to update omics data researchers about current data practices adopted by EGA based on 1) FAIR (Findable, Accessible, Interoperable, and Reusable) data principles, 2) implementation of Beacon networks as an open-source protocol for making anonymised genomic data discoverable for research and clinical purposes, and 3) implementation of technical standards such as Data Use Ontology (DUO) and authorization and authentication infrastructure (AAI) to ensure maximum utilization of data. Our aim with this abstract is to create collective effort in genomics data generation and accessibility

BioSeq14

Exploring the coding and non-coding miRNA targetome

Dimitra Karagkouni[1], Maria D Paraskevopoulou[2], Ioannis S Vlachos[2], Spyros Tastsoglou[1], Giorgos Skoufos[1], Artemis G Hatzigeorgiou[1]

[1]DIANA-Lab, University Of Thessaly, Volos, Greece, [2]University of Thessaly, Volos, Greece/Takeda Pharmaceuticals Inc., Boston, United States

microRNAs (miRNAs) are short (~23nts) non-coding RNAs, that act as central post-transcriptional gene expression regulators through target cleavage, degradation and/or translational suppression. More recently, miRNA:lncRNA (long non-coding RNA) interactions have been characterized.

DIANA-TarBase (http://www.microrna.gr/tarbase) is a reference database devoted to the indexing of experimentally-supported miRNA targets. Its 8th version is the first database to index >1 million entries, supported by more than 33 experimental methodologies, applied to 592 cell types/tissues under ~430 experimental conditions.

DIANA-LncBase (www.microrna.gr/LncBase) is a comprehensive repository of thousands of miRNA:lncRNA interactions supported by low/high-throughput, (in)direct experiments. The upcoming version of LncBase (October 2019) is significantly enhanced providing an unprecedented set of transcriptome-wide experimentally verified MREs on human and mouse lncRNAs on a wide range of tissues and cell types.

More than 60% of TarBase and LncBase content derives from the analysis of Argonaute crosslinking and immunoprecipitation (CLIP) experiments. Photoactivatable Ribonucleoside-Enhanced (PAR) CLIP methodology is considered one of the most powerful high-throughput methodologies for miRNA target identification. microCLIP (www.microrna.gr/microCLIP) is an innovative framework that combines deep learning classifiers under a super learning scheme for CLIP-Seq-guided detection of miRNA interactions. Former AGO-CLIP-guided implementations depend strongly on the T-to-C conversions to define miRNA bindings, while the efficacy of neglected interactions remained unknown. By analysing miRNA perturbation experiments and structural sequencing data, we showed that the previously neglected non-T-to-C clusters exhibit functional miRNA binding events and strong accessibility. microCLIP operates on every AGO-enriched cluster providing an average 14% increase in miRNA-target interactions per PAR-CLIP library, uncovering previously elusive regulatory events and miRNA-controlled pathways.

Indexing thousands of (non-)coding miRNA interactions is a valuable aid to the ncRNA community, demonstrated by the access of more than 6,000 users per month to the two databases content.

References

Paraskevopoulou MD and Karagkouni D, Vlachos IS, Tastsoglou S, Hatzigeorgiou AG, microCLIP super learning framework uncovers functional transcriptome-wide miRNA interactions, Nature Communications 9, 2018

Karagkouni D and Paraskevopoulou MD, Chatzopoulos S, Vlachos IS, Tastsoglou S, Kanellos I, Papadimitriou D, Kavakiotis I, Maniou S, Skoufos G, Vergoulis T, Dalamagas T, Hatzigeorgiou AG, DIANA-TarBase v8: a decade-long collection of experimentally supported miRNA-gene interactions, Nucleic Acids Res. 46, 2017

Paraskevopoulou MD, Vlachos IS, Karagkouni D, Georgakilas G, Kanellos I, Vergoulis T, Zagganas K, Tsanakas P, Floros E, Dalamagas T, Hatzigeorgiou AG, DIANA-LncBase v2: Indexing microRNA targets on non-coding transcripts, Nucleic Acids Res. 44, 2015

BioSeq15

High throughput sequencing uncovers patterns of UV damage formation

Elisheva Heilbrun[1], Hadar Golan Berman[2], Avital Wasserstrom[2], Sheera Adar[2]

[1]Jerusalem College Of Technology, Jerusalem, Israel, [2]The Hebrew University of Jerusalem, Israel

DNA damages are an obstacle to transcription and replication. Damages block a cell's ability to carry out its function and may lead to mutations and cell death. Ultraviolet (UV) radiation in sunlight is carcinogenic because it causes damages in DNA. The most abundant type of UV damages are pyrimidine dimers, primarily Cyclobutyl Pyrimidine dimers (CPDs), and pyrimidine (6-4) pyrimidone photoproducts [(6-4)PP]. These damages occur primarily in TT or TC di-nucleotides and are repaired by the nucleotide excision repair pathway. To study UV DNA damage formation and its repair, we employ state of the art methods that map the DNA damages and their repair at single nucleotide resolution across the human genome. Previous studies have demonstrated that UV dimers have no obvious damage hotspots. Damage frequencies are determined primarily by the underlying sequence composition. However, damage formation is not completely random but dictated by the different sequence composition of the different genomic elements. Our work uncovers transcription-associated asymmetries in damage formation that could affect UV-induced expression and mutagenesis.

BioSeq16

Quantitative Operating Principles of Yeast Metabolism during Adaptation to Heat Stress

Ester Vilaprinyo[1], Tania Pereira[2], Gemma Belli[1], Enric Herrero[1], Albert Sorribas[1], Rui Alves[1]

[1]Universitat de Lleida - IRBLleida, Lleida, Spain, [2]UCSF, United States

Microorganisms evolved adaptive responses to survive stressful challenges in ever-changing environments. Understanding the relationships between the physiological/metabolic adjustments allowing cellular stress adaptation and gene expression changes being used by organisms to achieve such adjustments may significantly impact our ability to understand and/or guide evolution. Here, we studied those relationships during adaptation to various stress challenges in Saccharomyces cerevisiae, focusing on heat stress responses. We combined dozens of independent experiments measuring whole-genome gene expression changes during stress responses with a simplified kinetic model of central metabolism. We identified alternative quantitative ranges for a set of physiological variables in the model (production of ATP, trehalose, NADH, etc.) that are specific for adaptation to either heat stress or desiccation/rehydration. Our approach is scalable to other adaptive responses and could assist in developing biotechnological applications to manipulate cells for medical, biotechnological, or synthetic biology purposes.

BioSeq17

Multi-Label Learning with Heterogeneous Label-Specific Features for Human Protein Subcellular Localization Sites Prediction

Hafida Bouziane[1], Abdallah Chouarfia[1]

[1]Université des Sciences et de la Technologie d'Oran Mohamed Boudiaf (USTO-MB), Oran, Algeria

Protein subcellular localization (SCL) is a crucial step to provide deep insights on protein function. To perform their proper cellular functions, proteins appear in specific compartments called organelles, where they reside or are in transit. Their presence in inappropriate compartments is responsible of many human diseases. Large-scale genome sequencing projects have generated a huge number of uncharacterized proteins which are still under intensive research. Nowadays, great efforts are invested for the annotation and characterization of these proteins using both experimental and in-silico techniques. The latest are still required to perform this challenging task, due to their low cost and reasonable prediction accuracy. Various predictive models have been proposed to tackle the SCL problem for different species and different localization coverage, using different assumptions and strategies but no model appears to work consistently better on all species. Machine learning based protein SCL prediction is a typical case of imbalanced classification. However, besides the highly imbalanced distribution of the proteins in different locations, the second challenge remains their ability to appear in simultaneously two or more different locations. These two situations reduce the ability of the learning models to properly distinguish each associated location. Hence, it is essential and indispensable to conceive learning systems taking these situations into full consideration. So far, many efforts have been made inthis regard, predicting either single-localization or multi-localization proteins using sorting signals, protein sequence/structure based features and gene ontology functional domains annotation which has shown a great interest. Many features-based methods used features selection strategy to reduce the learning space retaining only the pertaining features of protein samples. To deal with such problem, in the present study, we considered SCL prediction task as a multi-label learning problem and try to label unannotated proteins from both protein sequence and known location information. However, a common practice to improve the generalization performance of multi-label learners consists in exploiting the labels dependency/correlation. In this paper, we propose a prediction model for human protein sequences based on the input space rather than the output spaceby using Label specIfic FeaTures (LIFT) approach based on k-means clustering algorithm and Support Vector Machine binary classification which exploits input space constructed from heterogeneous features collected from protein amino acid sequence such as Amino Acid Composition (AAC), Dipeptide Composition (DPC), Position Specific Scoring Matrix (PSSM) profiles and Gene Ontology (GO) terms to further improve the generalization capability. LIFT strategy infers the most pertinent and discriminative features for each class label to fed them into the learning model by performing two steps. Firstly, both negative and positive training instances associated to each class label are clustered to discern its specific features. Secondly, a family of classifiers are induced with each of them such asfor each class label, a new binary training set is created based on the constructed features. Our proposed model can predict 11 distinct locations in human proteins and can be easily extended to other organisms. Performance evaluation on Deeploc benchmark dataset showed competitive performance against state-of-the-art protein SCL prediction methods.

BioSeq18

Specific T-cell clones are associated with mammary tumor development

Hagit Philip[1], Miri Gordin[1], Sol Efroni[1], Alona Zilberberg[1], Moriah Gidoni[1], Raanan Margalit, Christopher Clouser[3], Kristofor Adams[3], Francois Vigneault[3], Irun R. Cohen[4], Gur Yaari[1]

[1]Bar-ilan University, Ramat-Gan, Israel, [2]Science in Action Ltd, Israel, [3]Juno Therapeutics, United States, [4]The Weizmann Institute of Science, Israel

Cancer immunotherapy by checkpoint blockade proves that an effective immune response to a tumor can be induced clinically. However, the specificity of the T cell response to cancer progression has been difficult to quantify, especially in the context of spontaneous tumor progression. To quantify this specificity, we here study the T-cell receptor (TCR) repertoires in mice spontaneously developing mammary tumors; we produced and studied 120 samples from the peripheral blood of FVB/NJ mice transgenic at the Erbb2 locus (all developing tumors) and from their non-transgene age-matched controls. We sequenced alpha and beta TCRs monthly for 8 months. By mining these data in tandem with human breast cancer samples from human breast cancer data from multiple sources, including sets of stages, immunotherapy trials, single-cell data and matched normal tissue, we now report that a small set of shared T cells clones dominates the repertoire's temporal behavior over tumor progression. These clones prove to be ubiquitous over multiple studies and are unique in their nucleotide sequence origin. The identity of these T cells may assist in therapeutics, their temporal dynamics may assist in diagnostics.

BioSeq19

Genomic analysis of an insect-microbial symbiosis on the summit of a Hawaiian volcano

Heather Stever[1], Gordon Bennett[1]

[1]*University of California Merced, Merced, United States*

Virtually all insects have intimate associations with microbial symbionts that either parasitize or provide adaptive traits to their hosts. Beneficial features include the synthesis of essential nutrients and tolerance of environmental extremes. To investigate the effects of microbial associations on the evolution, ecology, and biology of host insects, we are using next generation genomic sequencing technology and computational tools to characterize the genomic features and nutritional contributions of bacterial symbionts in endemic Hawaiian Nysius seed bugs (Hemiptera: Lygaeidae). While most Nysius are herbivorous and maintain obligate associations with a specific endocellular bacterial symbiont (Schneideria nysicola) in a specialized organ called a bacteriome, not much is known about the role of this bacteria in host biological processes. Furthermore, nothing is known about the role of microbes in the extreme dietary and habit shift of the wēkiu bug (Nysius wekiuicola), a flightless Nysius species endemic to the cinder cones at the 4,200 meter (elevation) summit of Hawaiʻi Island's Maunakea volcano. In this extraordinary habitat - on an island that is only approximately 500,000 years old - wēkiu bugs have uniquely adapted to tolerate desiccation, freezing temperatures, and intense solar radiation; and unlike their fully-winged herbivorous congeners at lower elevations, wēkiu bugs instead feed on the dead and dying insects that are deposited on Maunakea's summit by wind. Our preliminary results indicate that S. nysicola bacteria likely contribute several B-vitamins and essential amino acids to their herbivorous Nysius hosts. We also found that wēkiu bugs are lacking this symbiont and instead have an altered and much more diverse microbiome compared to other Nysius. We are continuing to use advanced genomic and computational methods to further understand the functional roles of S. nysicola in herbivorous Hawaiian Nysius, and to investigate how the loss of this symbiont, and the presence of other diverse microbes may have influenced the evolution, ecology, and biology of the highly-specialized wēkiu bug.

BioSeq20

Understanding the translation potential and evolution of cytoplasmic long non-coding RNA

Isabel Birds[1], Katerina Douka[1], David Westhead[1], Mary O'Connell[2], Julie Aspden[1]

[1]Univeristy Of Leeds, Leeds, United Kingdom, [2]University of Nottingham, United Kingdom

Long non-coding RNAs (lncRNAs) are transcripts of 200 nucleotides or more which are thought to be without coding potential. LncRNAs form a diverse class of transcripts and although they share features with protein coding mRNA, for the majority of these transcripts little is known about their function. Recent advances in ribosome profiling have led to the discovery that a subset of lncRNAs contain actively translated small open reading frames (smORFs) across a range of organisms (yeast, D. melanogaster, mouse and human). These translation events remain controversial. In particular, knowledge of the mechanisms which govern the recognition and translation of lncRNAs is limited.

We seek to understand the importance of these lncRNA translation events both in Drosophilids and mammals. We have identified 3 populations of both Drosophila melanogaster and human cytoplasm: a) Translated lncRNAs, b) Polysome associated non-translated lncRNAs, and c) Cytosolic lncRNAs. Translation has been measured using Poly-Ribo-Seq. To understand these different lncRNA populations we have explored coding potential, along with their sequences and structural motifs. We will examine the levels of conservation of these cytoplasmic lncRNAs across a range of species, with the aim of gaining further insight into the possible relationship between lncRNAs and the evolution of new protein coding genes. Preliminary analysis of cytoplasmic lncRNAs across the Drosophila genus, revealed a higher level of species conservation in the translated lncRNA population than in cytosolic lncRNA.

BioSeq21

Annotation of the Mysterious Germline-Restricted Chromosome in Zebra Finch (Taeniopygia guttata)

Kathryn Asalone[1], John Bracht[1]

[1]*American University, Washington D.C., United States*

Developmentally programmed genome rearrangements are rare in vertebrates but have been reported in scattered lineages including the zebra finch (Taeniopygia guttata). In the finch, a well-studied animal model for neuroendocrinology and vocal learning, one such programmed genome rearrangement involves a Germline-Restricted Chromosome, or GRC, which is found in germlines of both sexes but eliminated from mature sperm. Transmitted only through the oocyte, it displays uniparental female-driven inheritance, and early in embryonic development it is apparently eliminated from all somatic tissue in both sexes. The GRC comprises the longest finch chromosome at over 120 million basepairs and because the zebra finch genome project was sourced from male muscle (somatic) tissue a majority of the genomic sequence and protein-coding content of the GRC remains unknown. In 2018, we reported the first protein-coding gene from the GRC, which was found to be under positive selection pressures. Through computational methods, a more complete set of genomic data from the GRC was isolate, annotated, and analyzed. Here we report hundreds of new genes from the GRC, including genes enriched for spermatogenesis and steroid metabolic process.

BioSeq22

Whole genome sequencing to study the contribution of structural variation to human complex traits

Klaudia Walter[1], Brittany Howell[1], Eugene Gardner[1], John Danesh[2], Adam Butterworth[2], Matthew Hurles[1], Nicole Soranzo[1]

[1]*Wellcome Trust Sanger Institute, Cambridge, United Kingdom,* [2]*Department of Public Health and Primary Care, University of Cambridge, United Kingdom*

Structural variations (SVs) are changes in the human genome larger than 50bp that include deletions, insertions, duplications and inversions. SVs are responsible for the majority of nucleotide variation among human genomes, and they are also implicated in various diseases and in associated phenotypes such as cognitive disabilities, as well as in predispositions to obesity, cancer and other diseases.

We have generated whole genome sequence data (WGS) at 15X coverage using ~12,000 samples from the INTERVAL study (http://www.intervalstudy.org.uk). INTERVAL is a cohort study of approximately 50,000 blood donors, aged 18 years and older, who were consented and recruited from 25 National Health Service Blood and Transplant donor centres across England between 2012 and 2014. Participants are predominantly healthy, and completed online questionnaires about basic lifestyle and health-related information, including self-reported height and weight, ethnicity, current smoking status, alcohol consumption, doctor-diagnosed anaemia, use of medications (hormone replacement therapy, iron supplements) and menopausal status.

The WGS data allow calling of SVs, including copy number variations (CNVs) in coding as well as non-coding regions. We applied multiple calling algorithms to call different classes of SVs, including deletions, duplications, mobile element insertions (MEI), inversions and copy number variants (CNVs). After stringent quality control criteria and in-depth annotation, we evaluate the contribution of SVs and CNVs on blood cell phenotypes and a range of other phenotypes related to inflammation and immunity, including 92 plasma proteins and 995 metabolites. In particular, we assess the heritability of SVs that is not already explained by SNPs and INDELs, and investigate the burden of rare SVs on genes or exons, as well as their functional impact on the non-coding, regulatory part of the genome to study the genetic architecture of complex traits. The availability of WGS data at 15X coverage also allows detection of possible mosaicisms that could be linked to predisposition to disease.

BioSeq23

Importance of exon duplications in alternative splicing

Laura Martínez Gómez[1], Michael L. Tress[1]

[1]*Spanish National Cancer Research Centre (CNIO), Madrid, Spain*

Alternative splicing and gene duplication have been proposed as two of the major mechanisms providing protein functional diversity (1,2). From the point of view of the protein, there are essentially just two types of alternative splicing, indels (which can be further separated into insertions and deletions) and substitutions. Protein sequence substitutions can be distinguished by their position in the protein sequence (N-terminal, C-terminal, internal) or by whether or not they arose from tandem exon duplications.

Substitutions that arose by exon duplication, make up only a small proportion of the annotated substitutions in the human genome (3). However, studies in human, mouse and Drosophila have shown that alternative isoforms generated from homologous duplicated exons are significantly over-represented in mass spectrometry studies (4). Homologous exon splicing events have very subtle effects in terms of protein folding disruption compared to other splicing mechanisms and a number are implicated in development and disease (5). Despite this little is known about the biological relevance of most homologous exons.

Here we manually retrieved more than 250 pairs of homologous exons. We estimated the duplication dates based on sequence similarity searches in lamprey, fugu, zebrafish, spotted gar and coelacanth and manual curation using the Ensembl, UniProt, RefSeq and APPRIS databases (6-9). We found that almost 80% of the tandem duplications in the set were conserved all the way back to coelacanth (more than 10 times as many as other alternative exons) and detected peptides for more than 55% of the isoforms generated from these homologous exons (compared to fewer than 1% for all other types of splice events).

Our results suggest that the generation of alternative isoforms from exon duplications, while rare, is likely to be an important means of generating functional diversity in eukaryotes.

1 Ohno, S. 1970. Evolution by gene and genome duplication. Springer, Berlin

2 Mironov, A.A., Fickett, J.W., and Gelfand, M.S. 1999. Frequent alternative splicing of human genes. Genome Res. 9: 1288-1293.

3 Kondrashov, F.A.; Koonin, E.V. Origin of alternative splicing by tandem exon duplication. Hum. Mol. Genet. 2001, 10, 2661–2669.

4 Abascal F, Ezkurdia I, Rodriguez-Rivas J, Rodriguez JM, del Pozo A, et al. Alternatively Spliced Homologous Exons Have Ancient Origins and Are Highly Expressed at the Protein Level PLOS Computational Biology. 2015; 11(6): e1004325.

5 Hatje K, Rahman R, Vidal RO, et al. The landscape of human mutually exclusive splicing. Molecular Systems Biology. 2017;13(12):959.

6 Aken,B.L., Achuthan,P., Akanni,W., Amode,M.R., Bernsdorff,F., Bhai,J., Billis,K., Carvalho-Silva,D., Cummins,C., Clapham,P. et al.(2017) Ensembl 2017. Nucleic Acids Res., 45, D635–D642.

7 The UniProt Consortium. UniProt: the universal protein knowledgebase Nucleic Acids Res. 2017 ;45:D158–D169.

8 O'Leary,N.A., Wright,M.W., Brister,J.R., Ciufo,S., Haddad,D., McVeigh,R., Rajput,B., Robbertse,B., Smith-White,B., Ako-Adjei,D.et al. (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res., 44, D733–D745.

9 Rodriguez,J.M., Rodriguez-Rivas,J., Di Domenico,T., Vázquez,J., Valencia,A. and Tress,M.L. (2018) APPRIS 2017: principal isoforms for multiple gene sets. Nucleic Acids Res., 46, D213–D217.

BioSeq24

Meta-analysis of NGS data to study poised enhancers in ESCs along differentiation

Mar González-Ramírez[1], Enrique Blanco[1], Luciano Di Croce[1]

[1]*Center for Genomic Regulation, Barcelona, Spain*

Poised enhancers and bivalent promoters play a pivotal role in regulating multiple expression programmes during ESC differentiation. Poised enhancers are decorated by the repressive mark H3K27me3, while bivalent promoters are decorated by H3K27me3 together with the active mark H3K4me3. However, while bivalent promoters have been intensively studied, our knowledge on poised enhancers is very limited. Here I propose a novel computational approach to identify poised enhancers and to characterize them in terms of binding of Polycomb group of proteins, RNA Polymerase II, and other features. My approach consists in generating a chromatin state map using histone marks found in regulatory regions (H3K27me3, H3K27ac, H3K4me3, H3K4me1) to determine poised regions. Next, I performed the classification of these regions into bivalent promoters and poised enhancers. Moreover, the transition of poised enhancers into active enhancers during differentiation was also investigated. To achieve this purpose, I have analysed data at several time points of two differentiation processes: from ESC towards either cardiomyocytes or cortical neurons. I have identified which poised enhancers switch into an active state at each time point and which are their target bivalent genes. The results indicate that most bivalent promoters switch into an active state during both differentiation processes and their corresponding genes become expressed. However, among the bivalent genes, the lineage-specific ones are precisely those showing higher expression levels. From this analysis, I conclude that the tight activation of specific poised enhancers determines which lineage-specific bivalent genes should be highly expressed in order to achieve a proper differentiation process.

BioSeq25

Chlamydiales implication in the evolution of Archaeplastida

Marie Leleu[1], Mick Van Vlierberghe[2], Ugo Cenci[3], Steven Ball[3], Denis Baurain[2]

[1]Lille University / Liège University, Liège, Belgium, [2]InBioS-PhytoSYSTEMS-Phylogenomics of Eukaryota, Belgium, [3]UGSF - UMR8576, France

Within Eukaryota, many lineages have emerged as the result of endosymbiosis. Hence, the unique association between a heterotrophic unicellular eukaryote and a cyanobacterium has led to a new organelle called the plastid and responsible for the spread of photosynthesis to eukaryotes. This major evolutionary event, also known as the plastidial primary endosymbiosis, has given birth to the supergroup Archaeplastida, composed of Rhodophyta (red algae), Glaucophyta and Chloroplastida (green algae and plants). Recently, a paradigm shift in the acquisition of photosynthesis has proposed the implication of an intracellular obligate pathogen in the primary plastid establishment. This hypothesis, dubbed the Menage-a-trois Hypothesis (MATH), suggests an active and direct role of Chlamydiales in the cyanobacterium-heterotrophic eukaryote primary endosymbiosis, which would have provided many critical genes to the cyanobiont in the common inclusion vesicle. The expression and efficient localization of these genes, such as key transporters and glucan transferases, would have initiated the biochemical fluxes of symbiosis. Even if still controversial, the MATH is supported by molecular, biochemical and phylogenetic evidence. Hence, studies performed more than a decade ago concluded that 30 to 100 genes would have been transferred from Chlamydiales pathogens to the ancestor of Archaeplastida. In this work, we revisit the phylogenetic support for the MATH with the objective of updating the list of Chlamydiales genes found in modern Archaeplastida and identifying a congruent phylogenomic signal (if any) corresponding to the potential original pathogen. Starting from all relevant publicly available data, we produced a representative set of primary algae and Chlamydiales genomes supplemented by non-photosynthetic eukaryotic and bacterial genomes. Selected proteomes were compared to each other and their proteins grouped into orthologous groups. Single-gene phylogenetic analyses then allowed us to automatically identify trees suggesting a Chlamydiales origin of the Archaeplastida proteins. In order to maximize the identification of the possible endosymbiotic gene transfers, this tree selection was based on the phylogenetic association between Chlamydiales and Chloroplastida and/or Rhodophyta, and between Chlamydiales and Glaucophyta. This way, we were able to identify around 200 genes (of which 30-40 had been already identified in previous studies) which may have been transferred from Chlamydiales to Archaeplastida during plastid establishment. Manual analyses are nevertheless necessary to confirm this number. A second round of phylogenomic analyses is ongoing, either based on explicit subsets of concatenated congruent gene alignments or using a Bayesian model (Bayesian Phylogenetics and Phylogeography software) that automatically clusters genes based on shared histories.

BioSeq31

Characterization of regulatory variants in promoters with enhancer activity and their relation with human diseases

Lucia Ramirez-Navarro[1], Salvatore Spicuglia[2], Lisa Strug[3], Jessica Denis[4], **Alejandra Eugenia Medina Rivera**[1]

[1]Universidad Nacional Autonoma De Mexico, Queratаro, Mexico, [2]Aix-Marseille Univ, INSERM UMR S 1090, Theory and Approaches of Genome Complexity (TAGC), F-13288 Marseille, France, [3]University of Toronto; The Hospital for Sick Children, Canada, [4]Vanderbilt Genetics Institute, Vanderbilt University Medical Center, Nashville, TN, United States

Gene regulation is driven by the interaction of regulatory sequences, commonly categorized as either enhancers or promoters. Recently, using a modification of the STARR-seq assay, we identified sets of promoters with enhancer potential. In a first publication the group characterized these promoters with enhancer activity (ePromoters), finding that these sequences share epigenetic characteristics with enhancers and do show contact with other promoters in 3D interactions. ePromoters represent between 2% to 6% of the promoters in the genome, and they show cell type specificity. Moreover, genes regulated by ePromoters show enrichment in gene ontologies related to inflammatory or stress response.

The majority of genetic variants associated with human diseases and traits (93.7%) have been found to be located in non-coding DNA, and particularly enriched in open chromatin regions, indicating that these variants may have an effect on regulatory mechanisms.

Using genetic variants associated with traits and disease (GWAS catalog), together with variant associated with gene expression in human tissues (Gtex eQTLs), we set out to identify relevant regulatory mechanisms that affect ePromoters function. Moreover, as associated SNPs and eQTLs are not necessarily causal variants, we identified the variants in linkage disequilibrium to them, extending the collection of variants of interest, overlaying this expanded collection to the ePromoters.

Using the GWAS catalog annotations to traits and diseases, we found a significant enrichment of GWAS variants associated to Hematological Measurements in HeLa ePromoters, while in K562 ePromoters additionally we see enrichment of “Other measurements” category, which tends to be related to different conditions as asthma or osteoarthritis, related to inflammatory response. In the case of tissue Gtex annotations, there were none significant annotations for single tissues showing enrichment.

In total 572 and 670 variants of our extended collection fall within the ePromoters neighbourhood in HeLA and K562, respectively, of these we identified 406 and 302 (K562 and HeLA, respectively) to be likely affecting binding for 11 transcription factors reported as enriched in ePromoter sequences. Particularly, we found the variant rs3771180 associated to asthma to be affecting the binding of the MAF::NFE2 and FLI1/FEV/ETS2/ELK4/ELK4/GABP1/Gabpa , which is an eQTL upstream the interleukin 1 receptor like 1 (IL1RL1) gene and in long range interaction with Interleukin18 receptor accessory protein (IL18RAP) gene. Supporting our hypothesis of ePromoters being associated to inflammatory response.

Understanding ePromoters and the regulatory mechanisms that affect their dual function will help identify the causes of human diseases and traits.

Compute06

GRASShopPER - a tool for de novo assembly

Aleksandra Swiercz[1], Wojciech Frohmberg[1], Artur Laskowski[1], Jan Badura[1], Marta Kasprzak[1], Jacek Blazewicz[1]

[1]*Institute of Computing Science, Poznan University of Technology, Poland*

Second generation sequencers produce billions of short DNA sequences in a massively parallel manner, which causes a great computational challenge in accurately reconstructing a genome sequence de novo using these short sequences. We propose GRASShopPER, GPU overlap GRaph ASSembler using Paired End Reads, which follows an approach of overlap-layout-consensus. It uses an efficient GPU implementation for the sequence alignment during the graph construction stage and a greedy hyper-heuristic algorithm at the fork detection stage. A two-part fork detection method allows us to identify repeated fragments of a genome and to reconstruct them without misassemblies. The method was tested on a benchmark data sets of a bacteria, nematode and human chromosome 14. The assemblies were evaluated with the golden standard tool QUAST. In comparison with other assemblers, GRASShopPER provided contigs that covered the largest part of the genomes and, at the same time, kept good values of other metrics, NG50 and misassembly rate.

Compute07

Prediction of small RNAs in bacteria and deciphering their interactions with proteins

Amita Barik[1]

[1] *National Institute of Technology, Durgapur, India*

Small non-coding RNAs (sRNAs) are reported to perform a myriad of molecular functions in bacteria. Identification and functional characterization of such RNAs hence, are important for understanding the sRNA-mediated regulatory networks in bacteria. Computational algorithms to predict sRNAs can expedite costly and time-consuming experimental methods to identify such RNAs. But unlike their eukaryotic counterpart, micro-RNAs, these RNAs are very heterogeneous and do not share a common pattern, making the computational detection extremely challenging. Nevertheless, many computational algorithms have been developed in the past for predicting sRNAs, but most of them are limited to a particular strain of bacteria with serious limitations and there is need for adjustments for their use on other genomes. Here, we propose to develop a computational method for predicting sRNAs in various bacterial species. We would be using both sequence- and structure- based features of experimentally validated sRNAs and based on the best features, a 10-fold cross validation on the training dataset will be carried out and a Random Forest (RF) model will be developed for differentiating sRNAs from the coding sequences. The designed algorithm would be made available as standalone program for users. Moreover, the interaction between sRNAs and RNA-binding proteins would be studied using bioinformatics approach which in turn will provide insights for role of sRNAs in post-transcriptional modifications.

Compute08

Enabling job scheduling flexibility in heterogeneous modular supercomputing systems

Ana Jokanovic[1], Julita Corbalan[2]

[1]Barcelona Supercomputing Center, Barcelona, Spain, [2]Universitat Politècnica de Catalunya, Spain

Today, modular supercomputing architectures are being deployed to meet the needs of applications from different scientific fields. Each module represents a cluster itself comprised of nodes with either CPUs or combination of CPUs and accelerators, GPUs and FPGAs. Applications are tuned for the best performance on the specific module or a combination of modules. However, the same application may run on any other module with suboptimal performance, i.e., longer time between start and end of the job. Still, the user needs shorter time-to-result, i.e., the shorter time between submission and end of the job. The flexibility of having preferred and alternative types of resources gives more options to the scheduler when deciding when and where to allocate the application. This might help the application to start and thus, to complete sooner, leading to better time-to-result than in one-module-request scenario.

In this work, we present the job scheduling policy in which allows the user to specify several modules to be considered for the execution of its application. First, we have analyzed the cases of six applications ranging from molecular dynamics to space weather simulations. The majority of applications can be executed on any of the several modules, but target specific modules for the best performance. They may run on different modules directly, or there exists a version for each module.

The scenario is supported in the SLURM's current version through the mechanism of partitions, but it is limited to homogeneous systems. Namely, the requested amount of resources and time are the same for any of the considered partitions. Thus, the data structures for the definition of user requests per module were not provided, and adapting the existing ones was too costly. To enable this flexibility in modular systems, we have extended the widely used job scheduler SLURM such that the user can specify the list of modules in the order of preference, adding a new sbatch option —module-list. The sbatch submits automatically one job to each of the modules specified in the module list. Also, we have implemented the model that converts the time and resource requirements from one module to another. This information is used to create the description for each of the module-list triggered job submissions. Finally, we have implemented a new dependency option plussingleton, which will be assigned to each of these jobs. The idea is that as soon as one job from the dependency group starts, all the rest are canceled. The policy aims to reduce the average slowdown, i.e., the ratio of the wait time and the execution time on the preferred module. We present the results obtained both from the real machine and from the SLURM simulator.

Compute09

Drug delivery simulations inside human body with a parallel high performance simulation platform, Alya

Ane Beatriz Eguzkitza[1], Guillaume Houzeaux[1], Mariano Vazquez[1], Constantine Butakoff[1]

[1]*Barcelona Supercomputing Center, Barcelona, Spain*

Flow simulations are used in the development of new cars and airplanes. The same technology can be used in humans. This is precisely the work presented here. We have been developed a complete fluid and particle dynamics model of respiration and cerebrospinal circuit.

Drug delivery simulations in the human body represent a non-invasive technique that is capable to open the black box that the human body is itself in most cases.

Due to the extreme complexity of these models, the use of large-scale computational resources together with efficient simulation codes is mandatory.
The development of computational patient-specific model deposition particle in the global respiratory system to study obstructive disease is presented.
Chronic Obstructive Pulmonary Disease (COPD) is a highly disabling airways pathology, with a high prevalence and a significant economic and social cost.
The prediction of the therapeutic responses thanks to computational models of pulmonary inflammation and their integration into clinical practice
is the main motivation of the INSPIRE project in which we have applied the work described here. To achieve this objective a sufficiently fine mesh is needed in order to capture all the scales present in the complex and turbulent flux which is responsive to transport the pollutants in the airways.
All of these ingredients will lead to a high fidelity simulation.
Another example of the application of transport particle simulations in airway trees is in the design of inhaler devices. The effect of aerosol therapies depends on the dose deposited beyond the oropharyngeal region as well as its distribution in the lungs. Factors such as the size of the aerosol particles, breathing conditions or the geometry of the airways play a fundamental role in the lung deposition of aerosolized drugs. These peculiarities of each individual make it necessary to have available in clinical practice a method to personalize aerosolized therapies.

Computational drug transport in cerebrospinal fluid (CSF) is a feasibility technique also presented in this work. Although the Reynolds number related to this flow is low, the complex and long geometry of the problem leads to huge meshes. The quality of the mesh is another key point in the success of the results and is not a trivial issue. Pulsatile flow with many cycles needs to be solved in order to stabilize the solution. Due to the really small size in some elements, the size of the time step also becomes really small and the resulting simulation involve huge computational resources. We have optimized the process but maintaining the precision of the results.
The simulations of the CSF in the spinal subarachnoid space for drug delivery studies can be a very useful strategy to improve the design of new catheters.

All of this work is developed with an in-house multi-physic parallel simulation code, Alya. This code, part of the European PRACE benchmark suite, is an HCP-based tool adapted to run efficiently on large-scale parallel computers. This involves Physical modeling, Mathematical algorithms, and code development and optimization, all with the strong constraint of efficient use of parallel resources.

Compute10

Condensed Microbiome representation using Transfer and Deep Learning to Promote Microbial Composition Prediction

Sara Cabello Pinedo[1], **Beatriz García-Jiménez[1]**, Mark D. Wilkinson[1]

[1]Center for Plant Biotechnology and Genomics UPM - INIA, Universidad Politecnica de Madrid, Spain

Motivation:
Data produced by metagenomic studies has multiple layers of complexity. Even 16S taxonomic analyses result in high-dimensional, extremely complex data that thwarts knowledge discovery. In this study, we describe a strategy to reduce the dimensionality of microbiome datasets, such that they can be interrogated and explored more easily.

Method:
This work brings together Deep Learning techniques, and microbiome data. We selected a particular type of artificial neural network - an autoencoder - to condense long vector values into a short vector (i.e. an encoded representation) which is more amenable to various kinds of analyses. In this case, the long vector of values describes a microbiome sample. We further show that we are able to recover the original vector from the encoded representation with high fidelity.

Results:
We transfer knowledge from a previously published dataset of around 5000 maize root microbiome samples into our autoencoder model, which returns a code of 6 rational numbers representing the information contained in the long vector of 717 taxa that describes the microbial composition of those samples. We are subsequently able to predict 458 of those taxa, after decoding, with a Pearson correlation greater than 0.5, with 0.77 being the average. This compressed representation opens-up many novel possibilities for microbiome data analysis, particularly with respect to knowledge retrieval and visualization. The autoencoder structure provides the ability to recover the complete abundance vector from the codified samples, making it possible to perform all analyses using the reduced coded data, and to recover the long vector only when required. For example, we apply our encoded microbiome to a novel scenario, showing that we are able to predict the final microbial composition (717 taxa, after recovery of the original vector) of maize root microbiome samples using only a few available environmental variables such as plant age, temperature or precipitation. We achieve an average mean square error of 0.0016; this is a higher accuracy than predictions made without our encoded model.

Conclusions and Further work:
This condensed representation could be applied to any environment (gut, ocean, urban soil, etc.) where there is a representative set of samples available. The contributions of our proposed microbiome autoencoder include: a) a novel dimensionality reduction approach to representing a long taxa vector as fewer than ten values; b) the ability to undertake challenging tasks in microbiome data analysis, such as to predict the microbial composition of hundreds of taxa based on a small number of features, rather than the more common (and simpler) task of predicting a phenotypic feature of the microbiome-associated host (e.g. age of the plant, productivity or disease) from hundreds of taxa; c) the encoded version of a microbiome can be reused, via transfer learning, into novel but related studies, allowing complex analyses to be undertaken using fewer de novo sequencing samples; the knowledge encoded within the microbiome autoencoder model can be applied to samples from a similar environment, enabling inferences or predictions in studies that would otherwise have insufficient power.

Compute11

Use of Machine Learning to Diagnose Inflammatory Bowel Disease using Associated Metagenomics Dataset

Hilal Hacılar[1], Ufuk Nalbantoğlu[2], **Burcu Bakir-Gungor[1]**

[1]Abdullah Gul University, Kayseri, Turkey, [2]Erciyes University, Turkey

Motivation:

Human gut microbiota is a complex community of microorganisms including trillions of bacteria that populate our gastrointestinal tract. While some of these microorganisms are considered as essential regulators of our immune system, some others can cause several diseases such as Inflammatory bowel diseases (IBD), diabetes, and cancer. In recent decades, the rapid advances in next generation sequencing technologies accelerate the discovery of the human microbiome. IBD, comprising Crohn's disease and ulcerative colitis, is a gut related disorder where the deviations from the "healthy" gut microbiome is considered to be associated with IBD. Since the etiology of IBD is not fully understood and its symptoms are complex, the design of new tools that make use of the available human gut metagenome data is essential for the diagnosis of IBD. In this respect, machine learning is well suited to obtain a diagnostic model using IBD-associated metagenomics dataset.

Method:

In this study, we aim to develop a classification model to aid IBD diagnosis and to discover IBD-associated biomarkers using metagenomics data. In this regard, the sequencing data of 148 IBD patients and 234 healthy individuals were fetched from MetaHit project and categorized into disease states based on the associated metadata. Sequencing reads were assigned to taxa using MetaPhlAn2 taxonomic classification tool. To deal with the high dimensionality of features, we applied robust feature selection algorithms such as Conditional Mutual Information Maximization (CMIM), Fast Correlation Based Filter (FCBF), min redundancy max relevance (mRMR) and Extreme Gradient Boosting (XGBoost). We compared the performances of machine learning methods such as support vector machines, k-nearest neighbor (kNN), random forest, adaboost, logitboost, decision tree and some hybrid methods. Additionally, we attempted to understand the underlying structure of IBD metagenomics data using Principal Component Analysis (PCA), and to find the subgroups of IBD patients using k-means and hierarchical clustering.

Results:

1455 taxa were identified from the IBD metagenomics dataset and used to train and test our model. In our experiments with 10-fold cross validation, we observed that XGBoost has a considerable effect in terms of minimizing the microbiota used for the diagnosis of IBD and thus reducing the cost and time. When we analyzed these identified species, we found that most of them are known as related with IBD development mechanisms. We also observed that compared to the single classifiers, ensemble methods such as kNN+logitboost resulted in better performance measures (95,6% AUC, 89% F1-score, 91,623% accuracy) for the classification of IBD.

Conclusions:

Metagenomic analysis of human microbiome reveals significant phenotypical signals, such as disease, as microbiome is modulated via human-microbiome symbiosis. Since the accuracy of diagnosis in IBD is key for a prompt and effective treatment, there is an utmost need to develop a classification technique which can expedite IBD diagnosis. In this respect, this study utilizes several supervised and unsupervised machine learning algorithms to increase the diagnostic accuracy of IBD, investigates potential pathobionts of IBD, and finds out which subset of microbiota is more informative than other taxa applying some of the state-of-the art feature selection methods.

Compute12

Mechanisms Underlying Allosteric Molecular Switches of Metabotropic Glutamate Receptor 5

Claudia Llinas Del Torrent[1]

[1] *Autonomous University of Barcelona (UAB), Spain*

The metabotropic glutamate 5 (mGlu5) receptor is a class C G protein-coupled receptor (GPCR) that is implicated in several CNS disorders making it a popular drug discovery target. Years of research have revealed allosteric mGlu5 ligands showing an unexpected complete switch in functional activity despite only small changes in their chemical structure, resulting in positive allosteric modulators (PAM) or negative allosteric modulators (NAM) for the same scaffold. Up to now, the origins of this effect are not understood, causing difficulties in a drug discovery context. In this work, experimental data was gathered and analysed alongside docking and Molecular Dynamics (MD) calculations for three sets of PAM and NAM pairs. The results consistently show the role of specific interactions formed between ligand substituents and amino acid side chains that block or promote local movements associated with receptor activation. The work provides an explanation for how such small structural changes lead to remarkable differences in functional activity. While this work can greatly help drug discovery programs avoid these switches, it also provides valuable insight into the mechanisms of class C GPCR allosteric activation. Furthermore, the approach shows the value of applying MD to understand functional activity in drug design programs, even for such close structural analogues.

Compute13

Evaluation of Natural Language Processing and Machine Learning modules for High-Performance Computing: Experiences with Tumor classification

Claudia Rosas[1], Joaquim Moré[1]

[1]Barcelona Supercomputing Center, Barcelona, Spain

Natural language processing may involve complex and costly processes that require not only the knowledge of the experts but a large number of resources in terms of time and computational capacity. Computational linguistics has proven to be a versatile field that can benefit any research area, being medicine and biology one of the greatest beneficiaries. To interpret written documentation enables the researchers to identify relevant terms from clinical histories or scientific evidence that leads them to draw significant conclusions or classify data more efficiently. There are widely known and tested tools and suites that have served to biology-related fields to facilitate these tasks, yet some of them require intermediate knowledge of programming languages such as Java or C++ to implement your processing modules.

Moreover, the bridge between the tools and learning algorithms to use the resulting data to extract further knowledge is not a smooth process. In this paper, we present our first exercise of classifying cancer tumors by using open source Python-based modules to process scientific evidence in the text and apply machine learning algorithms. The simplicity of this programming language reduces the time of prototyping. Our goal is to provide a leading example of a natural language processing pipeline that uses machine learning. Which not only can be used as a reference scenario when evaluating additional tools from different modules or platforms but can also be ported to high-performance computing machines to process data from this and other fields. Our experiences will serve as a starting point to reduce the gap between the use of this type of processes by the non-specialized public.

Compute14

How to Best Represent Target Proteins for Artificial Learning Based Drug Discovery and Repurposing

<u>Heval Ataş</u>[1], Ahmet Sureyya Rifaioglu[1], Rengul Atalay[1], Volkan Atalay[1], María Martín[2], Tunca Dogan[3]

[1]Middle East Technical University, Turkey, [2]EMBL-EBI, United Kingdom, [3]EMBL-EBI / Hacettepe University, United Kingdom

The identification of drug-target interactions (DTIs) constitutes the basis of computational drug discovery studies. To generate DTI prediction models using supervised machine learning techniques, input ligands and/or proteins are converted into quantitative feature vectors using various types of molecular descriptors for the training of the system. Therefore, the selection of descriptor sets is crucial to generate predictive models with high performance. While there are many studies for the benchmarking of compound descriptors (due to the abundance of ligand-based DTI prediction methods), protein descriptor analysis studies are scarce. In this study, we perform a benchmark analysis of various sequence-based protein descriptors considering not only physicochemical characteristics of amino acids, but also sequence composition, PSSM profiles and functional characteristics of proteins; using random forests (RF) and support vector machine (SVM) algorithms. The aim here is to identify the best representation of proteins to be used both in DTI prediction and for other types of automated protein annotation studies.

To investigate the protein descriptors, we assumed 2 different modelling approaches: (i) the target-based approach, in which an individual predictive model is generated for each compound cluster and the system is fed by only protein features; and (ii) the proteochemometric modelling (PCM) approach, in which both the compound-target feature pairs are fed to the system for the prediction of the actual binding affinities. PCM is a relatively new paradigm that utilizes both compound and protein space, which is critical to identify the interactions between compounds and proteins with low number of experimental data points. In the target-based approach, we generated nine independent bioactivity datasets from the ChEMBL database and used 42 different types of protein descriptors. In the PCM approach, we used ECFP4 fingerprints for the representation of compounds. For proteins, we selected 10 different descriptors out of the total 42, each taking a different aspect of proteins into account. For the training and test dataset, we used Davis kinase benchmark set by applying additional filtration steps for preventing bias and data memorization.

Although protein descriptors based on the physicochemical properties of a.a's are widely used for the representation of proteins, the results demonstrate that homology-based protein descriptors yield better performance for DTI prediction (i.e., k-sep as a PSSM-profile based descriptor provided consistently best performances in both analyses). Currently, we are performing this analysis on large-scale datasets of different protein families (i.e., membrane receptors, ion channels, transporters, transcription factors, epigenetic regulators and enzymes -with five subgroups-). These datasets are constructed from ChEMBL database by applying extensive filtering operations together with rigorous data partitioning strategies (e.g. scaffold and temporal based splits). Therefore, two significant outcomes will be acquired at the end of this study: (i) the best protein family specific sequence-based feature vectors, together with the discussion of results; and (ii) large-scale gold-standard datasets for different protein families, to be used as benchmark sets in DTI prediction studies, which is expected to fill an important gap in the field of predictive modelling for computational drug discovery and repurposing.

BioMed10

The Identification of Affected Pathway Subnetworks and Pathway Clusters in Colon Cancer

Miray Unlu-Yazici[1], Gokhan Goy[1], **Burcu Bakir-Gungor[1]**

[1] *Abdullah Gul University, Kayseri, Turkey*

Motivation:
Currently we encounter with omics revolution in which genome, epigenome, transcriptome, and other omes can be readily characterized. Traditional analyses attempted to untangle the molecular mechanisms of carcinogenesis using a single omic dataset which contributed towards the identification of cancer-specific mutations, epigenetic alterations, etc. However, the acquisition of cancer hallmarks requires molecular alterations at multiple levels. In this respect, combining multi-omic data is critical to understand the casual relationship between molecular signatures.

To facilitate the analysis of high-dimensional ‘omic’ data, protein-protein interaction (PPI) networks and biological pathways are widely used. Although several signaling pathways and driver genes crucial in the initiation and progression of cancer have been unveiled, a comprehensive picture of genomic and epigenomic changes in cancer development is far from being complete.

Method:
In this study, we proposed a novel method to detect affected pathway sub-networks and pathway clusters for a specific cancer type using multi-omics data. To test our method, transcriptomic and epigenomic data of colon adenocarcinoma (COAD) were obtained from TCGA. To analyze RNA-Seq and methylation data, edgeR Bioconductor package and ChAMP pipeline were used, respectively. The combined p-values of genes were calculated using Fisher’s combined test. Altered sub-networks which contain genes that are highly altered in transcriptomics and epigenomics datasets of colon cancer and are topologically close in the PPI network were identified using active sub-network search algorithms. Affected pathways were determined using hypergeometric test.

To create a pathway network, firstly, a binary gene-pathway matrix was generated. Secondly, a pathway-pathway matrix, which indicates the similarity between pathways was created from this gene-pathway matrix using kappa score. Then, affected pathway sub-networks and pathway clusters were identified on the generated pathway network.

Results:
COAD RNA-Seq data containing p-values of 9,426 genes and methylation data including 122,986 methylation changes of 18,299 genes were compiled for this study ($p<0.05$). When RNA-Seq and methylation data were combined, 7,546 common genes and their p-values were obtained. 690 affected sub-networks in a human PPI were identified and functionally enriched. A pathway sub-network including 93 pathways and 842 edges was identified from the generated pathway network with 288 KEGG pathway nodes and 10,904 derived connections. Within this pathway sub-network, 4 significant pathway clusters were obtained that included colorectal, pancreatic, endometrial, and prostate cancer, ErbB signaling, glioma, and melanoma pathways. Some of the affected genes within these identified pathways are already known in literature and we identified additional candidate genes.

Conclusions:
As the potency of single omics studies about the etiology of complex diseases is not sufficient, studies that combine multi-omics data and hence identify affected pathways have higher potential to discover cancer hallmarks. Since the pathways are strongly interrelated, here we proposed a novel approach that generates a pathway network using multi omics data, and identifies affected pathway sub-networks and pathway clusters. Our approach is based on both significance level of an affected pathway and its topological relationship with its neighbor pathways. The identified pathway sub-networks, pathway clusters and affected genes within these pathways helped us to illuminate colon cancer development mechanisms.

BioMed23

Model-driven discovery of metabolic reprogramming associated to metastatic cancer and cisplatin resistance

Marta Cascante Serratosa[1], Cristina Balcells[1], Carles Foguet[1], Miriam Tarrado[1], Oscar Camacho[1], Pedro de Atauri[1], Timothy Thomson[2], Francesc Mas[1], Silvia Marín[1]

[1]Universitat Barcelona, Barcelona, Spain, [2]Institut de Biologia Molecular de Barcelona-CSIC, Barcelona, Spain

Tumors harbor combinations of heterogeneous neoplastic cells. In this complex ecosystem, all modalities of mutual cell interactions can take place within the context of environmental cues that exert selective pressures. In spite of the underlying heterogeneity, two broad operational categories of neoplastic cells, namely cancer stem cell (CSC) and non-CSC, are most relevant with regards to two key properties of evolving tumor cells: survival to stress and metastatic colonization. Here, we apply a systems biology approach, including experimental data integration into genome-scale metabolic models, to unveil metabolic differences and potential vulnerabilities associated to metabolic heterogeneity of tumor cells subpopulations and to cisplatin resistance. We demonstrated that differential use of glucose and glutamine to fuel TCA cycle and mitochondrial respiration as well as differences at the level of central carbon metabolism could be exploited as putative drug targets in combined drug therapies.

Acknowledgments: MINECO-European Commission FEDER funds– “Una manera de hacer Europa”(SAF2017-89673-R), AGAUR–Generalitat de Catalunya (2017SGR1033) and Icrea Academia award 2015 granted to MC.

BioMed24

A Multi-Objective Genetic Algorithm to Find Active Modules in Multiplex Biological Networks

Elva María Novoa Del Toro[1], Efrén Mezura Montes[2], Matthieu Vignes[3], Élisabeth Remy[1], Diane Alexandra Frankel[1], Alexandre Atkinson[1], Annachiara De-Sandre-Giovannoli[1], Patrice Roll[1], Nicolas Lévy[1], Laurent Tichit[1], Anaïs Baudot[1]

[1]Aix-marseille Université, Marseille, France, [2]University of Veracruz, Artificial Intelligence Research Center, México, [3]Massey University, Institute of Fundamental Sciences, New Zealand

One of the most challenging tasks in computational biology is the integration of complementary biological data produced from different sources. We are particularly interested in the combination of expression data and biological interactions, with the objective to identify “active modules”, i.e., sets of interacting genes/proteins associated to expression changes in different biological contexts.

We developed a multi-objective genetic algorithm, capable of finding simultaneously several active subnetwork modules by considering both RNA-seq expression data and one or more biological interaction sources assembled in a multiplex network (e.g. protein-protein interactions, pathway interactions, and molecular complexes).

We use a modified version of the Non-dominated Sorting Genetic Algorithm II [1], with two objectives to maximize at the level of subnetworks: gene/protein differential expression and interactions’ density. The population of subnetworks is sorted via non-domination and separated into several Pareto fronts, depending on the values of these two objectives. We stop the search after a given number of generations and consider as a final result the set of non-dominated solutions. The main innovations of our algorithm are that i) it can retrieve several active subnetworks in a single run and ii) it can jointly leverage different biological networks, increasing the amount and type of information exploited.

We first tested our algorithm on simulated data. We defined artificially significantly differentially expressed subnetworks, as in Batra et al. [2]. We observed that our algorithm is able to locate most of the differentially expressed genes. Interestingly, the subnetworks we identify have an overall higher density as compared to the artificial subnetworks, although these latter ones have a higher level of deregulation. This comes from the fact that the selection of the initial subnetworks is not based on the density, calling for the development of more advanced benchmark strategies.

Then, we applied our algorithm to a real case study. We are interested in a very rare genetic disease, Hutchinson-Gilford Progeria Syndrome (HGPS), that is characterized by premature and accelerated aging. HGPS is caused by mutations in the LMNA gene that induces deregulations in nuclear shape and integrity, and associated cellular processes, such as gene expression, senescence, and apoptosis. We analyzed 8 RNA-seq fibroblast samples from 5 HGPS and 3 controls. We also built a multiplex network composed of three layers: a protein-protein interaction network (66,971 interactions), a pathways network (254,766 interactions) and a co-expression network (1,337,347 interactions) [3]. Our algorithm identifies 11 active modules from these data, which overall contain 116 genes and are enriched in different processes, from signaling to keratinization.

Benchmarking of multi-omics joint Dimensionality Reduction approaches for cancer studies

Laura Cantini[1], Pooya Zakeri[2], Aurelien Naldi[1], Denis Thieffry[1], Elisabeth Remy[2], Anaïs Baudot[2]

[1]IBENS, Paris, France, [2]Aix Marseille Univ, INSERM, MMG, CNRS, Marseille, France

Dimensionality Reduction (DR) approaches, decomposing datasets into low-dimensional spaces while preserving most of their original information content, are among the most prevalent machine learning techniques in data mining research. With the advent of high-throughput technologies, voluminous high-dimensional data have become a standard in biology, emphasizing the use of DR. This phenomenon is particularly pronounced in cancer biology, where national and international consortia have profiled thousands of patients for multiple molecular assays ("multi-omics"), including at the emerging single-cell scale.

Up to now, DR approaches have been mainly applied to the analysis of single omics data leading to cancer subtyping, tumor sub-clones quantification and tumor sample immune infiltration quantification. However, recently, advanced approaches designed to jointly integrate and analyze multiple omics have also been proposed. The various DR approaches are based on different mathematical assumptions, ranging from extensions of Canonical Correlation Analysis, tensors and more general data fusion approaches, which makes difficult to chose which method to apply.

We thus here propose an in-depth review and benchmarking of multi-omics DR approaches using:

i) artificial multi-omics data, corresponding to paired proteomic, transcriptomic and methylomic data generated starting from TCGA ovarian cancer data; in this case a ground truth on the clusters of samples is available and the methods can be compared based on the ability to retrieve it;
ii) multi-omics bulk data from 10 different cancer types downloaded from TCGA;
iii) multi-omics single-cell data from cancer cell lines.

In simulated data (i), the capability of the various methods to predict the clustering ground truth was found strongly sensible to the size of the clusters, with intNMF, RGCCA, MCIA and JIVE being the more robust methods. For cancer data (ii), MCIA, RGCCA, MOFA and JIVE more consistently identified factors associated to survival, clinical annotations and biological annotations. Finally, in (iii), tICA and MSFA outperformed other methods for their ability to cluster single cells based on their cancer cell line of origin. Interestingly, despite these good performances, tICA and MSFA have never been applied to single-cell data.

Overall the majority of the methods show convincing performances across all the three scenarios, with RGCCA, MCIA and JIVE, being the best performing ones. This suggests that a mathematical formulation, based on the search of omic-specific factors whose inter-dependence is maximized, better approximates the nature of multi-omics data.

The input data and all the analyses performed in this paper, going from the application of the various multi-omics DR approaches to the comparisons and figures generation, are fully reproducible through the use of a Jupyter notebook and a Docker image. This Jupyter notebook can be leveraged to test the DR algorithms on novel user-provided datasets, or to benchmark novel algorithms to the existing alternative approaches.

BioMed26

Targeting colorectal cancer: microbiome modulation and effect over tumour metabolism signalling pathways

Laura Judith Marcos Zambrano[1], Teresa Laguna[1], Enrique Carrillo[1]

[1]*Imdea Food Institute, Madrid, Spain*

Background:
Colorectal cancer (CRC) is the third most common cause of cancer worldwide, up to 10% of cases are hereditary, and the remaining 90% are sporadic. Risk factors for the development of sporadic CRC are widely described; recently, the intestinal microbiome has been pointed out as an essential factor influencing the onset and development of CRC. Moreover, it has been described changes related to the taxonomy and functionality of the microbiome according to CRC stages, and specific taxa/function present or absent according to the progression of the disease.
Modulation of the microbiome and the microbial metabolites implied in CRC progression would be of great interest. In this regard, precision nutrition interventions are of interest, considering that the microbiota could be modulated through diet and nutrients. Evidence suggests that dietary supplements could act as adjuvants in the treatment and management of cancer, but the possible modulatory effect over the microbiome of patients with CRC remains unexplored.

Objective:
We aim to study the metabolism signalling pathways interaction between microbiome and tumour cells in CRC, and the use of a bioactive compound derived from food phytochemicals to interfere with these pathways as an adjuvant treatment for CRC.

Research strategy:
Characterization of the microbial metabolome of patients with CRC based in public databases: We will evaluate previously published metagenomics datasets of oral (salivary) and gut (faecal) microbiomes to establish taxa composition, diversity, functional and metabolic changes related to CRC, taking into account the top-ranking features associated with CRC compared with healthy cohorts.

Determine the relationship between oral and gut metagenome and their interaction with CRC: We will select a cohort of patients with still untreated CRC admitted to Infanta Sofia Hospital, that fulfils the criteria for participating in the Clinical trial for the treatment with the bioactive compound selected. Samples of saliva and faeces will be obtained at the time of CRC diagnosis after the beginning of the treatment, and at the time treatment ends, before the surgical procedure for removing the tumour. Samples will be analyzed by shotgun metagenomics to characterize the oral and gut microbiome. We will analyze the taxonomic and the metabolic profile of the samples.

Evaluation of the modulatory effect of the bioactive compounds over the metabolic profile of the microbiota and the tumour progression: We will perform a metagenomics analysis of the samples obtained at the beginning and end of the treatment with the bioactive compound to evaluate shifts in microbiota function and composition. Moreover, we will assess the presence of the microbial biomarkers related to taxa and functionality described previously and monitor changes in those parameters after the treatment.

BioMed27

Applying human metabolic inter-variability for effective personalized nutrition strategies

Teresa Laguna[1], Laura Judith Marcos-Zambrano[1], Enrique Carrillo-de-Santa-Pau[1]

[1]*Imdea Food Institute, Madrid, Spain*

Currently precision nutrition has become of major interest, due to the outburst of personalized medicine. Similarly, the global aim of precision nutrition is to prevent and manage chronic diseases, particularly by adapting nutrition interventions and recommendations on the basis of individual characteristics. In cancer, alterations in glucose, nucleic acids and lipids metabolism have been largely known and cellular heterogeneity has been extensively explored. However, a global study of molecular heterogeneity of metabolism in cancer is missing, which will allow to identify target pathways for specific bioactive molecules in a personalized manner.

In this ongoing work, we analyze multi-omics data -transcriptomics and epigenomics- from healthy and tumor tissues to model the metabolism variability. Our approach is based on: 1) identify gene expression and DNA methylation variability in different healthy tissues, 2) analyze variability in tumoral tissues and identify altered metabolic pathways in several cancer types, 3) define distinct patient profiles corresponding to specific metabolic alterations in cancer, 4) explore target pathways of bioactives in order to improve the patients' health.

Preliminary analyses show that RNAseq count data from TCGA and GTEX databases display similar variability distribution in principal component analyses (PCA). The principal components of variability of the transcriptome data cover the inter-tissue heterogeneity. Thus, data from both consortia can be used subsequently in the study of transcriptomic and epigenomic variability. For that purpose, we will dissect the inter- and intra-tissue variability by calculating the coefficient of variation (CV) and the expression variability (EV) for each gene expression and DNA methylation data.

After modeling the metabolism variability and compare normal and tumor tissues heterogeneity, we will use public omics data from bioactive treatment studies to predict patients' response according to their metabolic profile. Finally, we expect to carry out clinical essays to confirm our findings.

BioMed28

Deciphering the interactions between the immune system and cancer cells

Victoria Ruiz Serra[1], Eduard Porta-Pardo[1], Alfonso Valencia[1]

[1]*Barcelona Supercomputing Center, Barcelona, Spain*

Every year, the worldwide incidence of cancer increases. Over 18 million people were diagnosed in 2018 alone and half of them are estimated to die from this disease. Part of this problem comes from the insufficient effectivity of standard treatments. Current research focuses on the development of strategies for treating those cancer types where other approaches fail. In this sense, the most promising next generation of cancer treatments are immunotherapies. Their advantage lies in the physiological cell-mediated immune response specificity when targeting tumor cells. So far, its effectiveness has been proven in cancer types where the standard treatments fail, such as metastatic lung cancer, metastatic melanoma and some blood cancers.

Although immunotherapies are a breakthrough in cancer medicine they are still limited to a subset of patients presenting certain clinical characteristics. Understanding the complexity of the intratumoral immune response will suppose a step closer to their complete success. Besides, we need a better insight on how the tumor-immune interaction network is affected by the accumulation of somatic cancer mutations.

Against the general tendency to study the effect of cancer-associated somatic mutations on a gene level, the present work pays attention to their impact on the immunological tumor microenvironment (TME) from a structural point of view. For instance, it is not the same to find a mutation affecting the catalytic site of a protein than in a non-functional area of the structure. We hypothesized that cancer-associated mutations involved in interaction areas or interfaces will provide a more comprehensive picture of why certain tumors have immune privilege. Our aim is to complement the gene-centered view of cancer.

To identify such interfaces, we analyzed more than 200.000 protein coordinate files from the Protein Data Bank. We defined protein interfaces as all the residues in close proximity (≤ 5Å) to either other proteins, nucleic acids or small ligands. We then used a linear model to describe the relationship between the location of missense somatic mutations from 10224 patients from The Cancer Genome Atlas and their tumor immune infiltration levels.

Our preliminary results show that 13379 of these interfaces, when somatically mutated, correlate with changes in the quantity of the immune infiltrate in the TME. Some hits correspond to typical cancer-associated genes such as CASP8, IDH1, or KRAS, where we are able to go beyond gene-centric analyses and distinguish different immune-effects depending on the region of the protein mutated. A less known example involves mutations on the interface between PAD1 and Ca2+ atoms, catalytic site of the enzyme. Patients with this interface mutated correlate with higher levels of immune infiltrate compare to mutations in other part of the protein or a wild type phenotype for that gene. Since PAD1 is in charge of protein citrullination, source of antigenic epitopes, perhaps a malfunctioning of its catalytic site could lead to the evasion of the immune system.

In summary, the analysis of individual protein interfaces provides valuable information to improve the understanding of the relationship between cancer cells and the TME.

BioMed29

Systematic integration of somatic mutation calling algorithms for reliable identification of cancer mutations

Marielena Georgaki[1], Panagiotis Moulos[1]

[1]Bsrc 'alexander Fleming', Vari, Greece

Problem: High-throughput sequencing has emerged as an effective strategy in research and clinical diagnosis in cancer. Many algorithms have been implemented for somatic single nucleotide variant (SNV) and insertion/deletion (INDEL) detection both in tumor-only and matched tumor-normal DNA samples. However, the resulting mutation profile can differ significantly among algorithms due to several non-trivial and unpredictable factors such as cross-contamination between tumor and matched normal samples and tumor heterogeneity. Moreover, algorithms that support somatic mutation calling without using matched normal sample have been proved inefficient by increasing the error rate to about 70%.

Method: Using Whole Exome, Whole Genome and Targeted Sequencing data, we compared the performance of publicly available somatic calling algorithms for matched tumor-normal and tumor-only samples. We focus on the detection of somatic SNVs and INDELs, derived from both simulated and real sequencing data, and the emergence of the state-of-the-art techniques that perform best in terms of accuracy and computational cost. Our goal is to provide the most reliable mutation profiles for cancer samples and for this purpose we are currently implementing a combinatorial algorithm for the integration of best performing individual variant detection algorithms.

The systematic integration is achieved by weighting mutation callers results according to their performance and by using Machine Learning approaches. The methodology also takes into account the interpretation of identified mutations which enhances their cancerous identity. To this end, we integrate important information derived from available and curated cancer databases such as CiViC, COSMIC.

Conclusion: As the first step in analyzing cancer sequencing data, detecting variants with high accuracy is of great importance in personalized medicine approaches. The systematic integration of reliable somatic variant callers that we propose, will enable researchers to capture somatic mutations more reliably and alleviate error rates derived from the multifunctionality and heterogeneity of a tumor.

BioMed30

Expanding the scope of drug repositioning in Juvenile Idiopathic Arthritis: A mechanistic approach using machine learning methodologies

Marina Esteban[1], Maria Peña Chilet[1], Carlos Loucera Muñecas[1], Joaquín Dopazo Blázquez[1]

[1]Clinical Bioinformatic Area, Sevilla, Spain

Introduction: Although pediatric rheumatology has seen a revolution in the therapies for rheumatic diseases, especially in juvenile idiopathic arthritis (JIA), current treatments are not sufficient to reach disease remission and to prevent long-term consequences. There is still a need to identify effective treatments that target specific steps of the immune response, to do so, it is crucial to comprehend the immunological mechanisms involved in the pathogenesis of the disease. Merging the knowledge about cell functional mechanisms and drug's action, together with the arising machine learning methodologies (ML), are resulting in the development of models capable of predicting quite accurately cell outcomes as well as the potential effect of external proteins over the disease map of action.

Material and Methods: We selected ORPHANET/OMIM (ORPHANET:618) well-known genes responsible for JIA, then, using KEGG DB, we constructed the disease map of action by extracting all the circuits (receptor-effector) containing those genes. We also selected well-characterized target genes of drugs approved by the Food and Drug Administration (FDA) or in late clinical trials. Using GTEx expression data, we obtained circuits activity and implemented Multi-Output Random Forest regression (backed by a repeated cross-validation strategy), to infer the effect of the selected drug-targets over the activity of the constructed disease map. The hyperparameters were selected by means of Tree-structured Parzen Estimator (TPE). In order to biologically validate resulting relevant drug-targets, we in silico simulated corresponding drug effects over available JIA samples on different datasets from GEO database (GSE21521 and GSE13501).

Results: The analysis of the implicated pathways in JIA, based on the 11 well-characterized genes involved in the disease, allowed the identification of 142 circuits that triggered disease pathogenesis and constitute the disease map of action. Moreover, the application of multi-output regression methodologies on the obtained disease map of action revealed relevant potential therapeutic targets. The evaluation of the in silico impact of relevant drug-targets can significantly accelerate the clinical translation of known compounds for novel therapeutic uses, opening a window for further investigations.

Conclusions: We have proposed a mechanistic approach, implementing ML methodologies, for the prediction of potentially causal relationships between drug targets and cell activities related to disease phenotypes, using the constructed map of the disease as a functional frame. The use of ML approaches may provide a new avenue to explore drug mechanisms, leading to a more efficient drug selection for patients treatment.

BioMed31

Exploration of unique chimeras as biomarkers in Alzheimer's disease

Orly Weissberg[1], Rajesh Detroja[1], Tomer Illouz[1], Alessandro Gorohovski[1], Dorith Raviv Shay[1], Eitan Okun[1], Milana Frenkel-Morgenstern[1]

[1] *Bar Ilan Universiy, Safed, Israel*

Dementia is a common disorder in the elderly age and is expected to affect 135 million individuals by 2050. The main cause of dementia is Alzheimer's disease (AD), which is characterized by accumulation of Amyloid-beta peptide, hyperphosphorylated tau protein and brain atrophy. AD patients have synaptic dysfunction and neuronal degeneration in several brain regions including the hippocampus, frontal cortex and amygdala. The disease is represented by progressive cognitive and emotional/behavioral impairments that eventually will lead to death. Merely 5% of AD cases are familial disease while the majority is sporadic. Until today only a few genes were tied with AD and only two drugs are being used to improve cognition during the early stages of clinical representation, but there is no available drug that can cure early or late onset of AD.

Liquid biopsy is non-invasive technique for detecting disease-related biomarkers in body fluids such as cerebrospinal fluid (CSF), plasma, serum or urine. We wish to study chimeras using this promising technique in order to find unique signature of AD related chimeras in CSF and plasma. To date, there is no unique biochemical test that can diagnose AD, excluding post mortem biopsies. Therefore, establishing Liquid Biopsy platform for AD will enable a simple and novel method to detect AD patient, hopefully in their early stages, and discover biomarkers that may be used as drug targets for treating the disease. We suggested to use chimeric transcripts produced by a slippage of two parental genes by means of unique exon-exon junction. To the best of our knowledge, chimeric transcripts of two independent genes that may be used as biomarkers have not been tested in AD yet. In our study, we found computationally 13 unique chimeric transcripts that involve parental genes known to be associated with AD. Moreover, other 20 novel chimeras were not previously associated with AD. Therefore, we attempt to validate the predicted chimeric transcripts in AD cortex biopsies and CSF, and then to use them as unique biomarkers and/or drug targets in non-invasive Liquid Biopsy platform for early diagnostics of AD.

BioMed32

Interpreting copy number variation pattern in cancer with protein interaction network

Qingyao Huang[1], Damian Szklarczyk[1], Michael Baudis[1], Christian von Mering[1]

[1]*University of Zurich, Switzerland*

Various types of genomic aberration are observed in malignant tumors, including single nucleotide mutation and copy number variations (CNV). While the former are focal changes, the effects of which can be localized to individual genetic elements, the latter often encompass metabases of genomic regions, covering hundreds of genetic elements. CNV has been successfully used on a coarse scale (cytoband level) as molecular features to define subtypes in many cancer entities. However, a systematic understanding of the frequent CNV patterns in cancers is not established. Difficulties include the functional redundancy, attenuation effect on transcriptomic and proteomic levels by feedback regulation. Therefore, knowledge from cellular network should be employed as a basis for such a genome-wide stoichiometric analysis. Here, we use STRING protein interaction network to establish a weighing method with network connectivity information on element-wise CNV data. With the METABRIC breast cancer data set and a set of breast cancer driver genes, we first establish that higher CNV coverage associate with poorer prognostic outcome, possibly due to late stage genome evolution. On top of this effect, we show that cancer samples with higher stoichiometic imbalance on driver genes show worse outcome. This study sheds light on the path to break down the evident CNV patterns as distinguishing features in cancer types and subtypes, and promote further application of genetic testing in cancer therapeutics.

BioMed33

Genomics tools in the cloud: The new frontier in omics data analysis

Rosa Barcelona[1], David Tomás[1], Andreu Paytuvi[1], Riccardo Aiese Cigliano[1], Walter Sanseverino[1]

[1]*Sequentia Biotech, Barcelona, Spain*

The knowledge and understanding acquired from genomics research can be applied in very diverse fields, including medicine, biotechnology and agriculture. Current sequencing platforms (Next Generation Sequencers) generate big data that need to be sorted, curated, integrated, analyzed and interpreted. Due to the size and complexity of such data, most geneticists or biologists fail to process their own data or it becomes time consuming. In addition, due to the diversity of genomic applications, informaticians usually do not have all the knowledge required to fully understand the biological “”problem”” of the researcher, making it impossible to reach a practical solution. Therefore, it is crucial to find a solution to efficiently manage the large amount of data produced as well as to analyze and interpret genomic data in an easy and efficient way.

Genomic variants obtained by sequencing already influence clinical decisions. They influence how physicians perform patients diagnose, prognostic evaluations and therapeutic choices. However, very few pipelines are versatile enough to call all types of genomic variations as they require very different algorithms, increasing the computational cost and highlight the need of a single platform, integrated and trustworthy.

On the basis of the above, we are developing an user-friendly platform that allow, in an accessible and scientifically robust way, to analyze and interpret genomics variations without need of informatic knowledge. For the time being, we have developed different pipelines for the analysis and identification of different types of variations (SNP, INDEL and CNV) as well as a parents-child trios analysis pipeline. Results are obtained in about 2,5 hours for 100x WES analysis (8 threads) and in 10 hours for 15x WGS analysis (8 threads). Interpretation and visualization of results is one of the bottleneck in these workflows, therefore, we have also built a visualization platform, which allows to display and browse dynamically in the results of SNPs/INDELs calls, and a database for data storage, which allow the user to have all samples stored in a unique environment without space problems. The visualization interface has two sections: variant browser and panels. Variant browser section focuses on the patient, allowing to visualize all the variants of a sample, whereas the panels section focuses on diseases, allowing to check whether a patient has known pathogenic variants related to any disease.

Our platform could introduce a disruptive innovation allowing researchers to perform analysis and integration of omic data easily, quickly and affordably. It is focused on professionals who wants to use genomics to perform decision making gathering genomics data with other evidences. For example, hospital professional who want upload their data and directly perform result interpretation, without giving rise to intermediate step as data analysis and processing. They could check directly the presence or absence of variations related to any disease of interest, enabling a quick decision-making approach. This is an innovative solution that aims to cover the gap between production and data analysis, specifically in the field of NGS data. The final result would be the democratization of significant part of the NGS bioinformatics.

BioMed34

Innate immunity receptors: computational approaches to their modulation

Sonsoles Martín-santamaría[1], Juan Guzmán-Caldentey[1], Alejandra Matamoros[1], Enrique Crisman[1]

[1]CIB-CSIC, Madrid, Spain

Toll-like receptor (TLRs) and lectins are pattern recognition receptors involved in the innate immunity. Particularly, TLR4 binds to lipopolysaccharides, a membrane constituent of Gram-negative bacteria, leading to the activation of the innate immune system response [1]. TLR4 activation has been associated with certain autoimmune diseases, noninfectious inflammatory disorders, and neuropathic pain, suggesting a wide range of possible clinical settings for application of TLR4 antagonists, while, TLR4 agonists would be useful as adjuvants in vaccine development and in cancer immunotherapy [2]. Among lectins, galectins specifically bind β-galactosides and play an important role in cell adhesion, signal transduction and cell recognition [3]. B-cell surface inhibitory protein CD22 (siglec-2) selectively recognize sialylated glycans, dampening autoimmune responses [4].

We have addressed the study of the TLR4 dimerization mechanism, and the influence of membrane composition, by building the full structure of the TLR4 complex inserted in the membrane with the help of coarse-grained and molecular dynamics simulations (unpublished results). Also, antagonist TLR4 modulators have been designed (glycolipids [5] and calixarene derivatives [6]) by combining docking, MD simulations and virtual screening, broadening the chemical spectrum for TLR4 therapeutics. Drug repurposing approaches have also allowed the identification of new TLR4 antagonists with non LPS-like chemical structure useful for the development of novel TLR4 modulators.

Regarding the modulation of lectins we have undertaken computational studies to deepen into the molecular recognition processes at atomic level. We have dissected the formation of CD22 homo-oligomers on the B-cell surface,[4] and studied ligand conformational entropy in galectin-3 recognition of histo blood-group antigens.[7] We are currently working on the design of improved selective modulators of galectins, by means of virtual screening, docking, and MD simulations, focusing on the computational exploration of adjacent pockets of the carbohydrate recognition domain. As result, we have obtained synthetic glycomimetics with improved affinity and selectivity (paper in progress).

References

[1] Park, B.S. et al. Nature. 2009, 458, 1191-1196.
[2] Kuzmich, N.N. et al. Vaccines. 2017, 5, 34, 1-25.
[3] Nabi, et al. Journal of Cell Science, 2015, 00, 1-7.
[4] Di Carluccio, C., et al. ChemBioChem, 2019. doi: 10.1002/cbic.201900295
[5] Cochet, F., et al. Sci Rep. 2019, 9, 919.
[6] Sestito, S.E., et al. J. Med. Chem. 2017, 60, 4882-4892.
[7] Gimeno, a., et al. Angew. Chem. Int. Ed. Engl. 2019, 58, 7268-7272.

BioMed35

Quantum methods for structure-based drug design

Ute Roehrig[1], Olivier Michielin[1], Vincent Zoete[1]

[1] *Sib Swiss Institute Of Bioinformatics, Lausanne, Switzerland*

One of the pillars of structure-based computational drug design is docking, which aims at predicting the binding mode of a ligand to a protein. Main challenges for docking algorithms include the treatment of protein flexibility, solvation, polarization, and covalent interactions. The last two problems are inherently quantum chemical in nature and can be addressed with our hybrid quantum mechanical/molecular mechanical (QM/MM) docking algorithm.

We will present results of our computational and experimental drug design tools applied to the design of indoleamine 2,3-dioxygenase 1 (IDO1) inhibitors. The heme enzyme IDO1 is involved in immunity, neuronal function and aging and serves as a target in different pathologies such as cancer and neurodegenerative diseases. Our fragment-based computational drug-design effort led to the discovery of highly efficient IDO1 inhibitors, the most active being of nanomolar potency both in enzymatic and cellular assays, while showing no cellular toxicity and a high selectivity for IDO1 over related enzymes.

BioMed36

Non-Structural Protein 1 (NS1) – A Hub Protein Essential for Influenza Infection – Using a Molecular Dynamics Approach to Understand its Behavior

Nícia Ferreira, Rita Melo[2], Carlos Barreto[3], Irina Moreira[3], Rui M. M. Brito[1]

[1]Chemistry Department, Faculty of Sciences and Technology, University of Coimbra, Portugal, [2]Center for Nuclear Sciences and Technologies, Instituto Superior Técnico, University of Lisbon, Portugal, [3]Center for Neuroscience and Cell Biology, UC – Biotech Parque Tecnológico de Cantanhede, Portugal

Around 650.000 people die from Influenza infections every year and 10% of the world population is infected. Influenza (flu) attacks mainly the respiratory tract being responsible for seasonal infections or, more rarely, pandemic outbreaks [1]. Currently, the better prevention method is an annual vaccine that has a limited efficacy. The lower rates of efficacy correlate to high rates of mutation and recurrent genetic assortment. New approaches to both prevent and treat Influenza are in need and only an improved understanding of the virus-host interactions can help in this demand [2].

From the 11 proteins encoded by the Influenza virus, non-structural protein 1 (NS1), due to it's great involvement in decreasing the host's innate immune response and increasing the viral replication rate, comes forward as a new interesting target for new therapeutics approaches [3]. NS1 is a 2 domain protein connected by a linker. The linker that connects the N-terminal RNA-Binding Domain (RBD) and the Effector Domain (ED) is a short, flexible region without a defined/fixed number of residues[4]. With 26 kDa size and being able to form homodimers, this protein is pivotal for the viral infection with its quaternary conformational statuses that shift depending on its partners, location in cell, and linker size [5].

This biological system presents itself as a highly evolutionary-conserved protein, but there is lack of information about its structure and behavior [6]. To this end, we performed six replicas of 1 µs Molecular Dynamics (MD) simulation of each NS1 system in order to fully characterize the conformational space visited by both domains and how the linker length and mutation on a pivotal residue influence the overall ability to establish protein-protein interactions. In particularly, we carried out MDs of 4 different full-length NS1 proteins: i) Wild-Type (WT) and a 15 amino-acids linker; ii) WT and a 10 amino-acids linker ; iii) W187Y NS1 mutant with a 15 amino-acid linker; and iv) W187Y NS1 mutant with a 10 amino-acid linker. This mutation is known to lower the prevalence of dimers in solution. Overall, our approach offers a new strategy for a better understanding of the NS1 homodimer dynamical behavior in human cells.

References:

1. Iuliano, A.D., et al., Estimates of global seasonal influenza-associated respiratory mortality: a modelling study. The Lancet, 2018. 391(10127): p. 1285-1300.

2. Mubareka, S. and P. Palese, Influenza Virus: The Biology of a Changing Virus, R. Rino and G.D. Giudice, Editors. 2011, Springer, Basel.

3. Das, K., et al., Structural basis for suppression of a host antiviral response by influenza A virus. Proceedings of the National Academy of Sciences, 2008. 105(35): p. 13093-13098.

4. Engel, D.A., The influenza virus NS1 protein as a therapeutic target. Antiviral Res, 2013. 99(3): p. 409-416.

5. Carrillo, B., et al., The influenza A virus protein NS1 displays structural polymorphism. J Virol, 2014. 88(8): p. 4113-4122.

6. Kleinpeter, A.B., et al., Structural analyses reveal the mechanism of inhibition of influenza virus NS1 by two antiviral compounds. J Biol Chem, 2018.

BioSeq30

Selected aspects of essential hypertension and cardiovascular disease - modeled and analyzed using Petri nets

Agnieszka Rybarczyk[1], Marcin Radom[1], Dorota Formanowicz[2], Piotr Formanowicz[1]

[1]Institute of Computing Science, Poznan University of Technology; Institute of Bioorganic Chemistry, PAS, Poznan, Poland, [2]Department of Clinical Biochemistry and Laboratory Medicine, Poznan University of Medical Sciences, Poznan, Poland

Essential hypertension is the world's most prevalent cardiovascular disorder, however, its etiology is still poorly understood, what makes it difficult to study. Recent studies suggest that inflammation can lead to the development of this type of hypertension and that oxidative stress and endothelial dysfunction are involved in the inflammatory cascade. In this work, to verify the influence of the selected factors on the development of the studied phenomenon, a Petri net based model has been built and analyzed. The analysis of the model has been based mainly on t-invariants and in silico knockout experiments.

This has enabled for an in-depth analysis of the studied phenomenon and has led to valuable biological conclusions. It has been shown that the most significant impact on the essential hypertension development has the activation of the RAS (the renin angiotensin system). It affects among others: the formation of angiotensin II, inflammatory properties (by influencing CRP), initiation of blood coagulation, activation of lymphocytes in hypertension, activation of NADPH oxidase, which is a key enzyme of an oxidative stress.

BioSeq32

Evaluation of Single-Molecule Long-Read Sequencing Technologies for Structural Variant Detection in Human Genomes

Nazeefa Fatima[1], Adam Ameur[1]

[1]SciLifeLab, Uppsala University, Uppsala, Sweden

Chromosomes can undergo various changes such as large deletions and/or insertions, resulting in structural variation differences between individuals. Structural variants (SVs) are a common source of variability in the human genome and are known to be associated with several diseases. SVs often involve complex genomic rearrangements that are difficult to resolve using short read sequencing technologies. New approaches enabled by the latest generation of long-read single-molecule sequencing instruments, provided by Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT), can produce a sufficient amount of data to enable SV detection across entire human genomes to a reasonable cost.

Previously, we performed PacBio sequencing of two Swedish human genomes, as part of the SweGen 1000 Genomes project (https://swefreq.nbis.se) and uncovered over 17,000 SVs per individual (Ameur et al, 2018). A majority of these SVs were not detectable in short reads. As a follow-up, we have now generated data for the same individuals on ONT's PromethION system, a new nanopore-based platform known for its higher throughput as compared to PacBio.

We present a pilot study that evaluates nanopore data derived from whole-genome sequencing (WGS) on PromethION in comparison to the Single-Molecule Real-Time (SMRT) reads obtained from the PacBio RS II platform. We performed comparative analyses of single-molecule technologies in a context of mappability, and SV detection that resulted in an average of 17k and 24k variants across nanopore and SMRT datasets, respectively. The results will be useful for the large-scale SweGen project, while the study serves as a bioinformatics pipeline for future long-read data analyses and sets a basis for what to consider when designing future PromethION experiments.

BioSeq33

The whole genome de novo sequence assembly of Nepenthes khasiana, a rare and endemic tropical pitcher plant of Meghalaya, North-east India

Ruchishree Konhar[1], Debasis Dash[1], Devendra Biswal[2]

[1]CSIR-Institute Of Genomics and Integrative Biology, New Delhi, India, [2]Bioinformatics Center, North-Eastern Hill University, Shillong, India

Introduction: Carnivorous plants have emerged as model systems for addressing a wide range of ecological and evolutionary questions. Nepenthes khasiana, an endangered carnivorous plant endemic to North-east India has been listed in Appendix I of CITES and is protected by the Wild Life (Protection) Act 1972 of India. In the present study, we report the de novo whole genome assembly of the Indian pitcher plant for the first time using NGS to investigate the molecular causes behind botanical carnivory in the monotypic family Nepenthaceae of the Order Caryophyllales.

Methods: High quality gDNA was isolated from the N. khasiana leaf sample using Qiagen Genomic-tip 100/G; quality and quantity ensured by Agarose gel electrophoresis and NanoDrop/Qubit Fluorometer. The shotgun and mate pair libraries were sequenced on NextSeq 500 using 2 x 150 bp chemistry. The chloroplast (cp) genome was assembled using adapter trimmed shotgun reads in NOVOplasty. For de novo whole genome assembly, adapter and quality trimmed shotgun and mate-pair reads were assembled using AllPathsLG. GapCloser and RepeatMasker was used for closing gaps and masking repeats. Repeat masked draft genome was used for all further analysis. Gene prediction was done with AUGUSTUS using Arabidopsis as the training dataset. SSRs were identified with MISA.

Results: 88.6% of the coding genome after final assembly has been found to be complete based on core orthologs (plants ortholog dataset). A total of 7,214 scaffolds were assembled with scaffold N50 1,163,181 bp (~1Mb) and average scaffold size 120 Kb. The genome size is computed as 749,857,876 bp(~750 Mb) based on k-mer distribution. The draft genome is more rich in tri-nucleotide repeats as compared to the mono- or di-nucleotide repeats. Assembled cp genome is 156,914 bp long with a quadripartite structure (a pair of inverted repeats, a large single copy and a small single copy region including 87 PCGS, 37 tRNAs and 8 rRNAs). N. khasiana whole genome data accession in SRA: SRP149035; Cp genome accession in GenBank: MH923233.

Discussion: The high quality reads from six paired-end libraries and three mate-pair libraries of N. khasiana was generated and assembled. The final assembly obtained by filling the gaps resulted in genome of 869.8 MB size. Identifying and masking repeat elements generated the draft genome with a total of 63,792 protein-coding genes. 5,067 genes were annotated into 24 functional pathway categories based on KEGG pathway database. 26,368 in silico validated SSRs have been identified in N. khasiana pitcher plant. Gene Ontology via BLAST2go annotation reveal enzymes involved in insect tissue digestion (chitinases, proteases and lipases) that might have evolved through evolutionary modifications of genes already existing in most angiosperms. Comparative genomics carried out between the predicted protein sequences of N. khasiana and protein sequences of Arabidopsis thaliana, Theobroma cacao, Vitis vinifera, Cephalotus follicularis and Populus trichocarpa led to the identification of 9,079 commonly occurring protein clusters.

BioSeq34

Next generation sequencing (NGS)-based de novo assembly of expressed transcripts and genome information of Dendrobium nobile, an endangered medicinal orchid from North-east India

Ruchishree Konhar[1], Devendra Biswal[2], Pramod Tandon[2]

[1]CSIR-Institute Of Genomics and Integrative Biology, New Delhi, India, [2]North Eastern Hill University, India

Introduction: The medicinal orchid genus Dendrobium belonging to the Orchidaceae family is the largest genera comprising about 800-1500 species. D. nobile is one such species with high medicinal value that demonstrate antiviral and anticancerous activity. With the availability of NGS technologies, D. nobile genome and transcriptome is achieved in the present study that has a potential to identify important pathways for the production of special and unique compounds of medicinal value.

Methods: Plant samples were collected from NRCO, Sikkim, India. The whole genome, total transcriptome and small RNA sequence data were all successfully completed using NGS technology. The progress and outcomes achieved out of this work is outlined in Fig.1.

Results: Assembly statistics of the whole genome, transcriptome assemblies are shown in Tables 1 & 2. Of a total of 180364 proteins, 77501 were annotated. Gene ontology chart and annotated pathways based on the KAAS analysis and gene distribution are shown in Figs. 2 & 3 respectively. Comparisons of the gene expression across different parts of orchid (leaf, stem, flower, root) are outlined in Fig. 4. Comparison of the DEGs reveals a large number of genes expressed at high levels in the flower and then in stem. Overall, based on hierarchical clustering, root has the most distinct transcriptome profile while leaf and stem have similar gene expression profiles.

Discussion: Orchid Multi-Omics Profiling Expression Database (OMOPED) (Fig. 5) is developed which integrates omics data from different sources and is powered by Biomart framework developed by EMBL-EBI. NGS data from the genome, transcriptome and small RNA sequencing for D. nobile as well as publicly available orchid datasets have been integrated for comparison. Users will be able to perform the following task with dedicated interfaces

- Browse data
- ID conversion between orchid orthologs
- Sequence retrieval
- Query Expression and compare orchid datasets
- Enrichment analysis

BioSeq35

De-novo protein function prediction using DNA binding and RNA binding proteins as a test case

Sapir Peled[1], Olga Leiderman[1], Rotem Charar[1], Gilat Efroni[1], aron Shav-Tal[1], Yanay Ofran[1]

[1]*Bar-Ilan University, Ramat-gan, Israel*

Proteins are the key players in cellular function. Of the currently identified protein sequences, 99.9% have never been observed in the laboratory as proteins and their molecular function has not been established experimentally.

Predicting the function of such proteins relies mostly on annotated homologs, under the assumption that two homologues share the same function. However, this has resulted in some erroneous annotations, and many proteins have no annotated homologs or no homologs at all. An additional set of method relies on a solved structure of the protein, which is not available for most proteins. A function prediction method that does not rely on any previous knowledge of the protein, can provide valuable insights regarding protein function.

Here, we propose a de-novo function prediction approach based on identifying biophysical features that underlie function. DNA and RNA binding proteins, essential proteins that handle our genetic information, were chosen as our test case. Using our approach, we discover DNA and RNA binding proteins that cannot be identified based on homology and validate these predictions experimentally.

This sequence-based function prediction tool, named DR. PIP (DNA/RNA Protein Interaction Predictor), uses the Random Forest algorithm for machine learning, and has two levels of prediction. The first one is predicting nucleic acid binding per residue, and then, combining these results- predicting if the protein as a whole can bind nucleic acid. We assess the performance of Dr. PIP computationally using large sets of proteins and show that it successfully predicts the function of DNA and RNA binding proteins directly from sequence and that it correctly identifies the site to which the nucleic acids bind.

We also use DR. PIP on a dataset of proteins with no known homologs, and discover DNA and RNA binding proteins that cannot be identifies based on homology. These predictions are validated using the B1H System. The performance of DR. PIP on these proteins was compared to six other online DNA binding prediction tools. DR. PIP was the only prediction method that was able to separate between the positive and negative samples, with higher scores to DNA binding proteins.

One such example is members of the Fibroblast Growth Factors Family, a family of mostly secreted proteins that is not known to bind DNA. FGF14 was predicted to bind DNA by DR. PIP. This prediction was verified experimentally. Moreover, the prediction of the binding site was also test experimentally. Mutating the predicted binding site on FGF14 abrogated DNA binding.

We also show that FGF14 is localized to the nucleus, supporting our predictions that it binds DNA.

DR. PIP is an efficient tool for sequence based de-novo prediction of DNA and RNA binding proteins, and for annotation the binding site on the protein. These findings show that an automated de-novo function prediction using biophysical features is possible. Such prediction can lead us to easier and cheaper annotation of protein function. Understanding protein function is a key to understanding biological processes and the way life operates.

BioSeq36

Deciphering the Influence of the Exome on Mutations

Sarah Wooller[1], Graeme Benstead-Hume[1], Frances Pearl[1]

[1]*University Of Sussex, Brighton, United Kingdom*

Improving our understanding of DNA structures and patterns, and their impact on mutations, is allowing us to better understand which areas of the genome are most at risk from different types of mutations, and the mechanism of damage.

Here we analyse the small mutations from over 32,000 exome-wide sequences in the COSMIC database.

We show that the nucleotides found at the site of short indels have more influence than those at the site of substitution mutations.

We find that somatic indels are predominantly single nucleotide deletions (and less frequently insertions) arising at the site of mononucleotide repeats between 1 and 7 nucleotides long.

Very few samples have lots of indels and a predominance of insertions and we find evidence that this is associated with the loss of MSH3 together with mutational signature 3.

There is an excess of in-frame indels in the coding region that cannot be explained by replication slippage. It is likely that this reflects the lower ability of cells to cope with highly pathogenic frameshift indels than inframe indels, giving rise to negative selection.

BioSeq37

Alignment-free evolutionary events prediction

Yulia Suvorova[1], Eugene V. Korotkov[1]

[1]*Research Centre of Biotechnology RAS, Moscow, Russian Federation*

Modern protein-coding sequences were formed by different evolutionary events. One of them is a fusion - a combination of two previously independent sequences or their parts into one new gene. Most of the current bioinformatics methods for the prediction of the fusion events are based on the sequence similarity. These methods utilize BLAST and it's analogs to search for individual subsequences which formed the supposed fusion in databases, in order to find possible ancestral independent sequences. However, if the ancestral sequences were lost during evolution or changed too much, it is impossible to detect the fusion.

Previously, it was shown that most protein-coding sequences have a triplet periodicity (TP). TP is absent in the non-coding sequences and introns. TP persists in the presence of substitutions, so this property needs more time to be changed [1]. We have developed a method for TP change points detection in protein-coding sequences and showed the relation of this phenomenon with the gene formation as a result of fusion [1, 2]. Our method detects fusion events on a statistically significant level (2-3% of the probability of Type I error). The method was applied to simulated fusions and real protein-coding sequences from eukaryotic genomes. Further analysis showed that about 30% of the TP change points can be explained by amino acid repeats. Another 30% can be potentially fused genes, alignment for which was found by the BLAST program. We believe that the rest of the results can be fused genes, the ancestral sequences for which have been lost [3].

Another important type of evolutionary events which can be detected based on TP is frameshift mutations. It has been shown that TP is associated with the active reading frame of a gene, and frameshift results in a TP phase shift. Here we developed a method that determines the best TP matrix for a sequence taking into account the correlation of the adjacent nucleotides along the sequence with the possibility of indels on each position [4]. The genetic algorithm and dynamic programming methods were used to identify the best TP matrix. After identifying the best matrix, we performed the final alignment of the sequence with respect to the matrix to find the probable frameshift positions. This method does not require any preliminary training or additional data. Using the developed method, coding sequences from the Arabidopsis thaliana genome were analyzed. In total, the algorithm found 9,930 (21%) sequences containing one or more potential reading frameshifts.

1. Suvorova, Y.M., Rudenko V.M, and Korotkov E.V. “”Detection change points of triplet periodicity of gene.”” Gene 491.1 (2012): 58-64.
2. Suvorova, Y.M., Korotkova M.A., and Korotkov E.V. “Comparative analysis of periodicity search methods in DNA sequences” Comput. Biol. Chem 53 (2014): 43-48.
3. Suvorova, Y M., and Korotkov E.V. “”New method for potential fusions detection in protein-coding sequences.”” J. Comput. Biol. (2019).
4. Suvorova, Y. M., et al. “”Search for potential reading frameshifts in cds from Arabidopsis thaliana and other genomes.”” DNA Research (2019).

BioSeq38

The positive part of phylogenetic varieties

Marina Garrote-lópez[1], Marta Casanellas[1], Jesús Fernández-Sánchez[1]

[1]*Universitat Politècnica De Catalunya, Barcelona, Spain*

It is well known that Phylogenetics and Mathematics, in particular Algebraic Geometry, are closely related. It is common to model evolution adopting a parametric statistical model which allows to define a joint probability distribution at the leaves of phylogenetic trees. When these models are algebraic, one is able to deduce polynomial relationships between these probabilities, and the study of these polynomials and the geometry of the algebraic varieties that arise from them can be used to reconstruct phylogenetic trees. However, not all points in these varieties are biologically relevant. In this talk, we would like to discuss the importance of studying the subset of these varieties with biological sense and explore the extent to which restricting to these subsets can provide insight into existent methods of phylogenetic reconstruction. One of our main focuses is to understand and describe these subsets of points that come from positive parameters. We are interested in the algebraic and semi-algebraic conditions that describe them and in knowing which of these conditions are relevant for topology inference. We will show some results on trees evolving under groups-based models and, in particular, we will explore the long branch attraction phenomena.

BioSeq39

CRISPR/Cas9 in silico off-targets prediction for Mucopolysaccharidosis Type I through comparative analysis

Martiela Vaz De Freitas[1], Paola Carneiro[1], Ursula Matte[1]

[1]*Federal University Of Rio Grande Do Sul, Porto Alegre, Brazil*

The latest genomic-editing technique, CRISPR/Cas9, is characterized by a single-guided RNA (sgRNA) with about 20 nucleotides. This sgRNA is complementary to a particular genomic region and directs Cas9 nucleases to introduce double-strand breaks in this region. Mucopolysaccharidosis type I (MPS I) is an autosomal recessive disease related to the deposition of glycosaminoglycans due to the deficiency of the lysosomal enzyme alpha-L-iduronidase (EC 3.2.1.76) encoded by the IDUA gene. Among the variants for the disease, p.Trp402* is the most frequent in patients with MPS I, in different populations. Genetic editing using the CRISPR system allows the development of a new therapeutic alternative for patients with MPS I. However, sequences similar to the target site may be also targeted by the CRISPR/Cas9 system. However, sequences similar to the target site may also be targeted by the CRISPR system. In this context, the objective of this work is to evaluate, in silico, potential targets for the CRISPR/Cas9 system. For this, five public software was used to evaluate potential off-targets sequences similar to the target site that make them potential cleavage sites. Preference was given to off-target sequences without mismatches and/or indels in the 5bp adjacent to PAM sequence, since that region is closely related to the activity of Cas9 at the target site. A total of 63 sequences were obtained as potential cleavage regions in the human genome beyond the target site. From these, 21 sequences have up to 6 mismatches and no indels and some containing both. Among these results, five off-target sequences were chosen for structural evaluation at mismatch positions, through homology modelling using PDB structures as template containing Cas9 protein, sgRNA, and a target DNA. The resulting predicted targets will be validated by the assembly of a gene panel for next generation sequencing of in vitro gene edited human fibroblasts from patients with MPS I.

BioSeq40

RSAT: Regulatory Sequence Analysis Tools

Nga Thi Thuy Nguyen[2], Bruno Contreras-Moreira[3], Jaime Castro-Mondragon[4], Walter Santana-Garcia[2], Raul Ossio[5], Carla Daniela Robles-Espinoza[5], Pierre Vincens[2], Denis Thieffry[2], Alejandra Medina Rivera[2], Jacques Van Helden[6], **Morgane Thomas-Chollier[1]**

[1]IBENS, Paris, France, [2]Institut de Biologie de l'Ecole Normale Superieure (IBENS) – INSERM 1024 CNRS 8197 – 46 rue d'Ulm 75005 Paris, France, [3]Estacion Experimental de Aula Dei-CSIC – Zaragoza, Spain, [4]Centre for Molecular Medicine Norway (NCMM), Nordic EMBL Partnership, University of Oslo, Norway, [5]Laboratorio Internacional de Investigacion sobre el Genoma Humano (LIIG) – UNAM, Santiago de Queretaro, Mexico, [6]Theory and Approaches of Genome Complexity (TAGC) – INSERM UMRS 1090, Aix-Marseille Université – Marseille, France

RSAT (Regulatory Sequence Analysis Tools) is a suite of modular tools for the detection and the analysis of cis-regulatory elements in genome sequences. Its main applications are (i) motif discovery, including from genome-wide datasets like ChIP-seq/ATAC-seq, (ii) motif scanning, (iii) motif analysis (quality assessment, comparisons and clustering), (iv) analysis of regulatory variations, (v) comparative genomics. Six public servers jointly support 10 000 genomes from all kingdoms. The latest novel or refactored programs include updated programs to analyse regulatory variants (retrieve-variation-seq, variation-scan, convert-variations), along with tools to extract sequences from a list of coordinates (retrieve-seq-bed), to select motifs from motif collections (retrieve-matrix), and to extract orthologs based on Ensembl Compara (get-orthologs-compara). RSAT is a long-standing, well-documented resource, available through Web sites, SOAP/WSDL (Simple Object Access Protocol/Web Services Description Language) web services and stand-alone programs. To further provide interoperability with external resources, such as motif collections from JASPAR, a REST API is under development. The whole suite can be installed locally, including with virtual machines, and a conda installation is under development to further ease local installation.

Availability : http: //www.rsat.eu/.

References :

- Nguyen NTT*, Contreras-Moreira B*, Castro-Mondragon JA, Santana-Garcia W, Ossio R, Robles-Espinoza CD, Bahin M, Collombet S, Vincens P, Thieffry D, van Helden J#, Medina-Rivera A#, Thomas-Chollier M#. “”RSAT 2018: regulatory sequence analysis tools 20th anniversary””, Nucleic Acid Research, 46(W1):W209-W214 (2018) https://doi.org/10.1093/nar/gky317

- Castro-Mondragon JA, Jaeger S, Thieffry D, Thomas-Chollier M#, van Helden J#. “”RSAT matrix-clustering: dynamic exploration and redundancy reduction of transcription factor binding motif collections.””, Nucleic Acid Research, 45:13 e119 (2017) https://doi.org/10.1093/nar/gkx314 Nga Thi Thuy Nguyen

BioSeq41

Fragment-based protein design guided by evolutionary relationships

Noelia Ferruz[1], Francisco Lobos[1], Dominik Lemm[2], Steffen Schmidt[1], Birte Höcker[1]

[1]Universitat Bayreuth, Bayreuth, Germany, [2]Universitat Pompeu Fabra, Spain

Nature has generated an impressive universe of proteins via replication, recombination, and differentiation from a set of protein fragments that serve as building blocks. While the complete de-novo design of proteins has proven a challenging task, the design of proteins via fragment recombination has met a broader range of success. We previously identified a conserved fragment between the TIM-barrel and the flavodoxin-like superfold [1] and used it to design a novel chimeric protein via heterologous recombination[2]. Here, wondering how many of such fragments exist and what structures they present, we extended our fragment search to the full protein universe. To this end, we took the SCOP2.06 protein database and performed all-against-all profile hidden Markov model comparisons of the domain sequences. This analysis identified hitherto unrecognized homologous protein segments between domains belonging to different folds. We later structurally superposed the region corresponding to these fragments and stored these data in our F(old P)uzzle database (fuzzle.uni-bayreuth.de).

Each entry in Fuzzle is defined by two protein domains that present a common fragment with a certain sequence and structural similarity. We analyzed high-scoring hits (p-value below 4.1× 10−5 and RMSD < 3.0 Å) using protein similarity networks, where the nodes represent domains that are linked if they have a fragment in common. Our analysis revealed that the protein universe presents a very large, continuous region populated by all-alpha and alpha/beta domains, which contains more than 50% of the network nodes, and discrete, island-like regions. We identified over 1000 protein components or fragments that span across the universe.

Besides the invaluable evolutionary information, these fragments can serve as building blocks for the design of new proteins. Based on the fragments stored in Fuzzle, we have implemented Protlego, an open-source, extensible, python platform that enables building protein chimeras between two proteins and permits its structural analysis and energy evaluation. As a result, a single, short Protlego script can lead from two PDB parent structures to a set of offspring chimeras, and compute useful properties such as hydrogen networks, hydrophobic clusters, contact orders, and potential energies. We believe that the identification and recombination of conserved protein fragments will open great opportunities for protein design.

References

[1] J. A. Farias-Rico, S. Schmidt, and B. Höcker, "Evolutionary relationship of two ancient protein superfolds," Nature Chem. Biol., vol. 10, no. 9, pp. 710–5, Sep. 2014.

[2] T. A. M. Bharat, S. Eisenbeis, K. Zeth, and B. Höcker, "A beta alpha-barrel built by the combination of fragments from different folds.," Proc. Natl. Acad. Sci. U. S. A., vol. 105, no. 29, pp. 9942–7, Jul. 2008.

BioSeq42

The anatomy of the human transcriptome in health and disease

Raquel Garcia-Perez[1], Mattia Bosio[1], Ferran Reverter[2], Miquel Calvo[2], François Aguet[3], Roderic Guigo[4], Pedro Ferreira[5], Kristin Ardlie[3], Marta Mele[1]

[1]Barcelona Supercomputing Center, Barcelona, Spain, [2]Universidad Autonoma de Barcelona, Spain, [3]Harvard University, United States, [4]Center of Genomic Regulation, Spain, [5]Institute of Molecular Pathology and Immunology, Portugal

Differences in gene expression determine phenotypic differences between individuals including differences in disease susceptibility. Most studies addressing how differences in gene expression are associated with phenotypic differences between individuals have focused on specific tissues or phenotypes in a limited number of individuals. Here, we take advantage of the GTEx v8 datasets which includes 14,891 transcriptomes from 838 individuals across 49 tissue types to understand how variation in gene expression and splicing is associated with over 30 human phenotypes including general phenotypes such as sex, ethnicity and age and, disease related phenotypes such as atherosclerosis, hypertension or BMI. Differential expression analysis were implemented using generalized linear models following the voom-limma pipeline, controlling for known and unknown batch effects (surrogate variable analysis). We find differentially expressed genes (DEG) across all 49 tissues for ethnicity, age and sex, with ethnicity showing the largest number of DEG in most tissues (~900 DEG) likely due to genetic effects. The number of age DEG is dramatically reduced once we control for tissue heterogeneity suggesting that the major contributor of age DEG are changes in cell-type composition. For the remaining phenotypes, we find that both DEG and differentially spliced genes (DSG) are restricted to a few tissues often including the tissue most related to the phenotype. We also observe unexpected associations such as a strong effect in the brain tissue in individuals with liver disease. Generally, the number of DSG is lower than the number of DEG for any phenotype with ethnicity being the largest contributor to differential splicing across most tissues (~273 DSG). We find little overlap (~15% on average) between DEG and DSG suggesting that both expression and splicing play a role in explaining to phenotypic differences between individuals. Overall, this study represents the largest transcriptome variation analysis to date and will provide deep insights into how phenotypic differences between individuals arise in health and disease.

BioSeq43

A benchmarking study of the performance of SNP callers for pool-seq data with an application for genomics analyses of multiple populations

Sara Guirao Rico[1], Josefa González[1]

[1]*Institut de Biologia Evolutiva, Barcelona, Spain*

The costs of genomic analyses of multiple populations, such as DrosRTEC and DrosEU initiatives, are still too high to address them through individual genome sequencing. In this case, pooling individuals for NGS can be a more effective strategy in SNP detection and allele frequency estimation because of a higher total coverage (mainly because the size of the samples is larger). However, it has been reported that SNP calling from pools are usually accompanied by high probabilities of sequencing error. Some approaches are intended to use a correction based on estimates of the sequencing error in subsequent analyses or a minimum threshold allele frequency required to include the SNP in posterior analyses. Nevertheless, these approaches depend on unbiased estimates of the error rate, generally unknown or difficult to obtain, or suffer from the inevitable loss of information on low-frequency variants, particularly relevant for population genomics analyses (both for demographic and selective inferences). Therefore, finding an optimal balance between minimizing information loss and reducing sequencing costs (i.e., increasing the effectiveness of allele frequency estimation) is essential to ensure the success of these ambitious projects. Here, we have developed a benchmarking study comparing the performance of most popular SNP callers for pool-seq data when estimating the site frequency spectrum under different realistic conditions. We are using both computer simulations (under a complex demographic history) and real data (the same samples sequenced separately for each individual and as a pool) to identify the best strategy to approach genomics studies of multiple populations under different conditions.

BioSeq44

A Survey on Different Approaches to Construct Phylogenetic Tree from SNP Data

Tazin Rahman[1], Faria Tabassum[1]

[1] *Washington State University, Pullman, Pullman, United States*

Phylogenetic analysis has been facilitated to a great extent due to the recent advances and the dramatic decrease in the cost of DNA sequencing technology. For evolutionary analysis of different species or organisms, phylogenetic trees are widely used. In recent times, single nucleotide polymorphism (SNP) data are broadly used in the phylogenetic analysis at deeper evolutionary timescales. SNPs are evolutionarily stable and work as biological markers. Researchers can use them to identify genes that are associated with diseases and track the inheritance of disease genes within individuals of families. The decreasing economical cost and the ease of collecting SNP data as a result of high-throughput sequencing have made them significant for inferring phylogenies. There are various approaches and tools to construct phylogenetic trees from SNP datasets. In this paper, we explore some tools and methods to construct phylogenetic trees based on SNP datasets. We also discuss the pros and cons of these approaches. The necessity of a standard and the simple automatic tool has become significant due to the importance of SNP data. The large expansion of genome sequence data calls for the requirement of carrying such analyses.

BioSeq45

Influence of function in bacteria concerning cancer

Tehila Atlan[1]

[1]Jerusalem College Of Technology, Jerusalem, Israel

Microbes are present everywhere: in the air, on the skin, within the colon and even in malignant tumors. These bacteria can cross-talk with all these different environments by producing enzymes and metabolites, which affect metabolic pathways of other living entities in their environment. Their influence may be beneficial, detrimental or insignificant to human health, depending on the context. For example, microbes living within the tumor microenvironments of some pancreatic tumors disturb chemotherapy protocols, by degrading the therapeutic agent, gemcitabine. Yet, despite this example, most microbiome studies focus on the phylogenetic aspects of the bacteria and do not yet attempt to evaluate the impact of the bacterial functions on the environment. Experiments based on the influence of the enzymes rather than the taxa of the microbes may be of huge interest since several different bacterial taxa stemming from different environments may actually converge to the same functionality. As proof of concept, we focus on lung tumors and focused on the different microbes and microbial functionalities between smokers and non-smokers. In the smokers' group, the functions were noticeably enriched with abilities to disassemble materials found specifically in cigarettes such as nicotine, tobacco-related components, xylene, aldehydes, methane, aromatic compounds. In most of the cases, we found more than a single kind of microbe in a specific sample and across samples that could potentially perform these types of metabolism. There is a wide variety of microbes in both groups (smokers and non-smokers) but they share very few of them. Our findings indicate that smoking greatly influences the lung microbiome of cancer patients and may have an additional carcinogenic affect via the microbial component of the tumor microenvironment.

BioSeq46

Chromatin 3D organization principles revealed by chromatin networks: gene-regulation, replication, and beyond

Emanuele Raineri[2], Miguel Madrid[1], **Vera Pancaldi[1]**

[1]Inserm Centre De Recherches En Cancerologie De Toulouse & Bsc, Toulouse, France, [2]CNAG-CRG, Spain

Recent technological advances have allowed us to map chromatin conformation and uncover the spatial organization of the genome inside the nucleus. These experiments have revealed the complexities of genome folding, characterized by the presence of loops and domains at different scales which can change across development and cell types.
Many approaches have been employed to describe 3D genome organization, which can be broadly divided into polymer physics models, constraint based models and statistical approaches.

An increasingly popular representation of chromatin is given by networks, in which genomic fragments are the nodes and connections represent experimentally observed spatial proximity of two genomically distant regions. This formalism, applied to promoter centred chromatin interaction networks generated by promoter capture HiC, has allowed us to consider a variety of chromatin features in association with the 3D structure. In particular, we exploited a known popular network metric to define Chromatin Assortativity: the tendency for regions of chromatin with similar properties to preferentially interact with each other. In addition to recapitulating known results, measuring chromatin assortativity of tens of features in mouse embryonic stem cells led us to novel biological insight on gene regulation [1].

Moreover, we have characterized DNA replication in a 3D chromatin context, generating novel maps of replication origins in mouse embryonic stem cells under normal conditions and during DNA replication stress. These origins were then contextualized by projection on a promoter-centred chromatin contact network defined at a few kb resolution. We found that replication origins with similar efficiency and genomic regions of similar replication timing interact with each other preferentially [2]. These findings suggest that DNA replication takes place in the context of hierarchical multi-scale structures spanning tens of megabases and even bridging chromosomes. More specifically, origins that interact with others tend to replicate earlier and with higher efficiency. The changes of origin activation patterns in normal and stressed conditions support a stochastic model of activation in which both local and global chromatin properties modulate efficiency.

Finally, we propose tools to investigate chromatin organization at different scales using networks, in particular an R package and an online chromatin network interaction viewer building on this framework. The ChAseR package allows users to efficiently integrate genome-wide datasets or lists of genomic regions with 3D chromatin interaction networks. It then efficiently computes Chromatin Assortativity of these features, highlighting the ones that are most strongly associated to genome architecture and performing different kinds of randomizations to assess the significance of these associations. Furthermore, we have developed GARDE-NET (https://pancaldi.bsc.es/garden-net), a web-portal where users can visualize multiple chromatin networks (>10 human PCHiC datasets and mouse embryonic stem cell PCHiC so far) in combination with pre-loaded chromatin features (histone modification peaks etc.) and with a chance to upload their own chromatin features of interest.

We will conclude by reflecting on general organization principles in genome architecture that can be revealed by applying the network formalism.

[1] Pancaldi et al. Genome Biology 17 (1), 152, 2016
[2] Jodkowska, Pancaldi et al. bioRxiv 644971, 2019

BioSeq47

Gene expression and splicing regulation in the colon helps to explain the genetic heritability of inflammatory and metabolic complex traits and diseases

Virginia Diez Obrero[1], Ferrán Moratalla[1], Robert Carreras-Torres[1], Víctor Moreno[1]

[1]Avoris Retail Division S.L., Palma de Mallorca, Spain

The functional role of genomic regions identified in large-scale genome-wide association studies (GWAS) is poorly understood and need to be thoroughly characterized. It is hypothesized that these loci are enriched by variants with tissue-specific regulatory roles and functionally relevant genes.

The aim of this study is to identify complex traits whose genetic heritability is partly explained by SNPs associated with changes in gene expression and alternative splicing in normal colon tissue.

We performed germline genome-wide genotyping and RNA sequencing on a novel collection of colon tissue biopsies from ~200 healthy individuals. We characterized their transcriptome by means of gene expression levels and profiled alternative splicing computing percent-spliced-in (PSI) index, which reflects the frequency at which specific splicing events occurs. Then, we identified cis-acting expression and splicing quantitative trait loci (eQTLs and sQTLs) and performed a genetic heritability enrichment analysis (LD score regression) over a list of 255 complex traits and diseases with GWAS summary statistics available that we considered being relevant for colon tissue.

Briefly, we found enrichment of genetic heritability for inflammatory bowel diseases (mainly Crohn's disease and ulcerative colitis), alcohol consumption frequency, glucose and insulin-related traits, circulating metabolic traits, body-max index, and other anthropometric traits. Overall, our findings provide evidences of the regulation of gene expression and alternative splicing in the colon tissue as potential underlying mechanisms of uncharacterized GWAS association signals.

Compute16

From tech to bench: Deep learning pipeline for image segmentation of High-Throughput High-Content microscopy data

Beatriz Garcia Santa Cruz[1], Javier Jarazo[2], Jens Christian Schwamborn[2], Frank Hertel[1], Andreas Husch[2]

[1]Center Hospitalier de Luxembourg / University of Luxembourg, [2]Luxembourg Center for Systems Biomedicine /University of Luxembourg

Automation of biological image analysis is essential to boost biomedical research. The study of complex diseases such as neurodegenerative diseases calls for big amounts of data to build models towards precision medicine. Such data acquisition is feasible in the context of high-throughput screening in which the quality of the results relays on the accuracy of image analysis. Although the state-of-the-art solutions for image segmentation employ deep learning approaches, the high-cost of manual data curation is hampering the real use in current biomedical research laboratories.

Here, we propose a pipeline that employs deep learning not only to conduct accurate segmentation but also to assist with the creation of high-quality datasets in a less time-consuming solution for the experts. Weakly-labelled datasets are becoming a common alternative as a starting point to develop real-world solutions. Traditional approaches based on classical multimedia signal processing were employed to generate a pipeline specifically optimized for the high-throughput screening images of iPSC fused with rosella biosensor. Such pipeline produced good segmentation results but with several inaccuracies. We employed the weakly-labelled masks produced in this pipeline to train a multiclass semantic segmentation CNN solution based on U-net architecture. Since a strong class imbalance was detected between the classes, we employed a class sensitive cost function: Dice coefficient. Next, we evaluated the accuracy between the weakly-labelled data and the trained network segmentation using double-blind tests conducted by experts in cell biology with experience in this type of images; as well as traditional metrics to evaluate the quality of the segmentation using manually curated segmentations by cell biology experts. In all the evaluations the prediction of the neural network overcomes the weakly-labelled data quality segmentation.

Another big handicap that complicates the use of deep learning solutions in wet lab environments is the lack of user-friendly tools for non-computational experts such as biologists. To complete our solution, we integrated the trained network on a GUI built on MATLAB environment with non-programming requirements for the user.
This integration allows conducting semantic segmentation of microscopy images in a few seconds. In addition, thanks to the patch-based approach it can be employed in images with different sizes. Finally, the human-experts can correct the potential inaccuracies of the prediction in a simple interactive way which can be easily stored and employed to re-train the network to improve its accuracy.

In conclusion, our solution focuses on two important bottlenecks to translate leading-edge technologies in computer vision to biomedical research: On one hand, the effortless obtention of high-quality datasets with expertise supervision taking advantage of the proven ability of our CNN solution to generalize from weakly-labelled inaccuracies. On the other hand, the ease of use provided by the GUI integration of our solution to both segment images and interact with the predicted output. Overall this approach looks promising for fast adaptability to new scenarios.

Compute17

Bioinspired Algorithms in Bioinformatics

Katya Rodriguez-Vazquez[1]

[1]IIMAS-UNAM, Mexico

In the field of Computer Sciences, Artificial Intelligence can be defined as the ability of reasoning of an artificial agent. In this area, it is used the term of “intelligent machine or intelligent computer”, but instead of “machines”, these are computer algorithms that automatically improve through experience. These are algorithms that “learn” and “adapt” to their environment; this means, a problem to be solved. Thus, we can talk about evolutionary and bio-inspired algorithms, techniques that try to partially replicate behavior of social cooperative and biological systems. Genetic algorithms, genetic programming, particles swarm optimization, ant colony systems, neural networks are classified as bioinspired techniques, among others. In this case, biological and social cooperative concepts are borrowed in order to propose computational algorithms. But, defining these bioinspired algorithms as starting point, these are applied to give solutions of diverse problems in the field of Biology, and now, Bioinformatics emerges as the area of computer sciences algorithms applied to the analysis and processing of biological data. In this talk, some examples will be presented as the sequences alignment (DNA, Proteins and metabolic pathways) problems using genetic algorithms, protein folding using a genetic algorithm, classification of microarrays by means of particle swarm optimization algorithm, and generation of prediction rules by means of genetic programming.

Compute18

Link-HD: a tool to multiple- microbial communities integration

María Laura Zingaretti[1], Gilles Renand[2], Diego Morgavi[2], Yuliaxis Ramayo[3]

[1]CRAG, Spain, [2]INRA, France, [3]INTA, IRTA, Spain

The drop in 'omics' technologies costs enables to obtain data from multiple data sources. However, the integration of these heterogeneous datasets is not a trivial task. Several statistical methods have been developed to handle these challenges. Here, we present a versatile tool to integrate multiple datasets. Our methodology is a generalization of STATIS-ACT ('Structuration des Tableaux A Trois Indices de la Statistique –Analyse Conjointe de Tableaux'), a family of methods designed to integrate information from multiple subspaces. Here, we extend the classical approach by incorporating distance matrices for numerical, categorical and compositional data; further, Link-HD performs variable selection using regression biplot, as well as, a differential abundance testing. Finally, our tool facilitates the interpretation of results through the taxon set enrichment and network analysis from selected variables. We illustrate the methodology integrating microbial communities from cows of which methane yield was measured. We also compare our results with MixKernel and we found a good concordance between common subspaces and features selected by both approaches. The source code, examples and a complete manual are freely available in https://github.com/lauzingaretti/LinkHD

Compute19

Deep and Shallow Chemogenomic Modelling for Compound-Target Binding Affinity Prediction Using Pairwise Input Neural Networks & Random Forests

Heval Ataş[1], Ahmet Sureyya Rifaioglu[1], Tunca Dogan[2], María Martín[3], Rengul Atalay[1], Volkan Atalay[1]

[1]Middle East Technical University, Turkey, [2]EMBL-EBI / Hacettepe University, United Kingdom, [3]EMBL-EBI, United Kingdom

The identification of binding affinity values between compounds and target proteins is critical for early stage drug discovery. Traditionally, binding affinity values are determined by high-throughput screening experiments, which are time-consuming and expensive, and thus, cannot be applied to the massive compound-target space. Therefore, computational methods have been developed to predict binding affinities, using machine learning (ML) techniques. Recently, chemogenomic modelling approaches became popular, where both compound and target protein features are used together as the input of the predictive models. Hence, they are able to incorporate targets with low number of (or no) training data and yield accurate predictions even for targets/compounds not involved in the training set at all. In this study, we developed two chemogenomics based receptor-ligand binding affinity prediction methods, using deep (pairwise input deep neural networks -PINNs-) and shallow (random forests -RFs-) supervised learning techniques, to predict the binding affinities of a large set of kinases against several drug candidate compounds.

We represented compounds with ECFP4 fingerprints, which is one of the most widely used feature type for compounds, and proteins with k-separated-bigram-PSSM feature vectors as a homology-based protein descriptor. The experimental bioactivity data points for kinases were obtained from the ChEMBL database for training (192,935 data points) by including all bioactivities containing a pChEMBL value (i.e., -log(IC50, EC50, Ki, Kd, Potency, ...)). For approach1, we used RF algorithm with tree number=100 and max_features=0.33. RF model takes a concatenated feature vector (compound + target) as input. For approach2, we used pairwise input feed-forward neural networks (PINN) as a deep-chemogenomic neural network architecture. The network takes a pair of feature vectors for compounds and targets from disjoint input nodes simultaneously, following 2 hidden processing layers, latent representation of compound and target features are concatenated and further processed on 2 additional hidden feed-forward layers. For both approaches, output is a single node (a regressor), which predicts binding affinity for the input compound-target pair in terms of pChEMBL values.

We participated the IDG-DREAM Drug-Kinase Binding Prediction Challenge (www.synapse.org/ drugkinasechallenge) with our models. The challenge is based on the prediction of 430 pKd values between 70 compounds and 199 kinases (round1) and 394 pKd values between 25 compounds and 207 kinases (round2). Considering the model performance results on challenge round1, our best performing model has reached an RMSE value of 1.119 (5th best team). In round2/final round, our RMSE performance was 1.066 (4th best team, overall).

The considerably high performance of our models in this challenge demonstrates the usefulness of chemogenomic approach for the computational prediction of compound-protein (i.e., receptor-ligand) interactions. This approach utilizes both compound and target space in one model, so that it can be used to predict novel ligands for targets with limited training data, and to identify the druggability potential of human proteins that were never targeted before. The approach developed in this study is expected to aid researchers in constructing high performance compound-target interaction predictors, especially for proteins with limited (or no) training data.

Compute20

MEXCOWalk: Mutual Exclusion and Coverage Based Random Walk to Identify Cancer Modules

Rafsan Ahmed[1], Ilyes Baali[1], Cesim Erten[1], Evis Hoxha[1], **Hilal Kazan[1]**

[1]Antalya Bilim University, Antalya, Turkey

Motivation: Genomic analyses from large cancer cohorts have revealed the mutational heterogeneity problem which hinders the identification of driver genes based only on mutation profiles. One way to tackle this problem is to incorporate the fact that genes act together in functional modules. The connectivity knowledge present in existing protein-protein interaction networks together with mutation frequencies of genes and the mutual exclusivity of cancer mutations can be utilized to increase the accuracy of identifying cancer driver modules. Results: We present a novel edge-weighted random walk-based approach that incorporates connectivity information in the form of protein-protein interactions, mutual exclusion, and coverage to identify cancer driver modules. MEXCOWalk outperforms several state-of-the-art computational methods on TCGA pancancer data in terms of recovering known cancer genes, providing modules that are capable of classifying normal and tumor samples, and that are enriched for mutations in specific cancer types. Furthermore, the risk scores determined with output modules can stratify patients into low-risk and high-risk groups in multiple cancer types. MEXCOwalk identifies modules containing both well-known cancer genes and putative cancer genes that are rarely mutated in the pan-cancer data. The data, the source code, and useful scripts are available at: https://github.com/abu-compbio/MEXCOwalk.

Compute21

Filling the gap between computing and I/O performance in Brain Tissue Simulations

Judit Planas[1], Felix Schuermann[1]

[1] *Ecole Polytechnique Fédérale de Lausanne, Switzerland*

In the past years, we have seen how the computing power has grown significantly, but the I/O performance has not increased at the same pace. This trend started several years before, but we only have global ranking lists for both concepts since 2017. If we take this period, since the beginning of the IO500 list [1] (2017-2019), and we compare it with the well-known Top500 list [2], we can see that while computing power has increased by 60% in Top500, I/O performance (bandwidth) has increased only by 18% in IO500. On top of all the technical challenges that this difference in growth implies, this translates into a very simple statement in computational science: "We can compute as much as we want, but we cannot process all the generated data". Therefore, computational scientists are struggling to manage and analyze all the data that their High-Performance Systems (HPC) systems are able to massively produce.

In this presentation we will stress the existing gap between computing and I/O capabilities and explain our attempts to fill in this gap, which fall in line with the general trends that international supercomputing facilities are following as well. More specifically, we have adopted a combination of software and hardware solution and worked close with the product vendor to adapt our scientific workflow. At the hardware level, this solution provides an extra I/O layer between the computing nodes and the file system, like a cache of the file system. It is based on SSD technology and is meant to accelerate the I/O bandwidth from the computing node point of view. At the software level, the commercial solution provides a library to manage the data that reside on that middle layer. With this product and some engineering effort on our side, we can run HPC simulations of brain tissues more efficiently.

Our presentation will focus on the engineering efforts on our side and our strategy to use the middle layer in two different ways: (i) to prefetch read-only data for the simulator; (ii) to cache simulation output data as it is being generated (and which usually accounts for 30% of the total runtime in a normal simulation). For the latter case, we have performed several tests at small scale with a prototype benchmark that mimics the simulator, showing promising results [3] and we are now in the process of adapting the real scientific workflow, so that it can take advantage of this new setup. In addition, we are also driving our engineering efforts towards a transparent integration, so that the scientists can use the simulator as they are used to and do not need to take the technical aspects into account.

[1] https://www.vi4io.org
[2] https://www.top500.org
[3] Ewart et al., Neuromapp: A Mini-application Framework to Improve Neural Simulators. Published in ISC 2017, DOI: 10.1007/978-3-319-58667-0_10

Compute22

Phage-Host interaction prediction

Laura Carolina Camelo Valera[1], Alejandro Reyes Muñoz[1]

[1]*Universidad De Los Andes, Bogota, Colombia*

Bacteriophages are known to strongly influence a wide variety of ecosystems, modelling ecological dynamics through the infection of bacterial cells and also promoting important evolutionary mechanisms such as horizontal gene transfer. However, due to advances in sequencing technologies and its applications in metagenomics, there exists a huge quantity of phage sequences lacking assigned hosts, leading scientists to misunderstand the role of phage-host interactions in many ecosystems, for instance, its effect in disease or health states. Recently, some authors have shown evidence of phage genome adaptation to its bacterial host genome in order to improve infection mechanisms, probably as a consequence of host-phage coevolution, meaning that nucleotide composition could be a relevant signal to infer the host of a given phage depending on its genome sequence.

Based on this assumption, this study aims to determine informative nucleotide signals for the host prediction of a given phage. To reach this objective, we trained a random forest model using kmer composition signals as features. First, we calculated k-mer usage bias for 3535 bacterial and 1148 dsDNA phage genomes from the order Caudovirales, then, we perform data balancing at the genus level for both the phage and bacterial datasets. Next, using the balanced datasets we build the random forest using all features, however, given the model variation and the signal to noise ratio, in the first place, we performed a feature filtering based on correlation and kmer similarity between phage and host genomes. Secondly, we implemented wrapper algorithms coupled with random forest to obtain the variables that were the most important to predict host taxonomic affiliation. Afterwards, we trained a random forest model on the selected features to create a method of phage-host affiliation prediction. Preliminary results have shown that training a random forest using the bacterial dataset and testing with the phage dataset gives a global classification accuracy around 40% for the bacterial host family comprising a good performance for the Pasteurellaceae, Helicobacteraceae, Burkholderiaceae and Lactobacillaceae families (F1 > 0.6), suggesting a space for model improvement in order to obtain a better method to predict the bacterial host.

Compute23

Classification model with Conformal Prediction applied to Glucocorticoid Receptor virtual library

M. Isabel Agea Lorente[1], Ivan Čmelo[1], Martin Šícho[1], David Sedlák[2], Petr Bartůnek[2], Daniel Svozil[1]

[1]Department of Informatics and Chemistry, University of Chemistry and Technology, Prague, Czechia, [2]Institute of Molecular Genetics. CZ-OPENSCREEN: National Infrastructure for Chemical Biology. Prague, Czechia

Defining the prediction boundaries of an in silico model is as important as determining its statistical quality. Training set objects define model's applicability domain (AD), which refers to the region of space where new predictions are considered to be reliable. The aim of this work is to develop the workflow for the systematic exploration of the chemical space of Glucocorticoid Receptor (GR) using a random forest classification model defining the AD with conformal prediction. GR plays significant roles in the pathology of serious human diseases, such as cancer or inflammation, and is, thus, widely used target in the field of drug design. In order to validate our model, we present a particular application of the classification model to a GR virtual library built using our molecular morphing algorithm (Molpher) [1] fed with active ligands from ChEMBL17 [2] and our in-house database. This method creates a path of molecules between a start and target point by the application of morphing operators that correspond to simple structural changes, like the addition or removal of an atom. If the start and target molecules are active at the same receptor, the molecules encountered along the morphing path, so called morphs, will represent a focused virtual library and actively enriched virtual libraries provide important information for further in silico and in vitro investigation. Prospective validation of Molpher against a new version of ChEMBL(ChEMBL24) revealed that the algorithm is able to generate new actives that are based on previously unknown scaffolds and that the model is able to correctly classify them.

Compute24

ElTetrado captures topological features of nucleic acid quadruplexes

Marta Szachniuk[1], Mariusz Popenda[2], Tomasz Zok[1]

[1]Poznan University of Technology, Poland, [2]Institute of Bioorganic Chemistry, Poland

Quadruplexes are unique tertiary structure motifs occurring in nucleic acid molecules. So far, they have been confirmed as promising therapeutic targets in many drug development strategies and contributors to various biological processes. Recent years, brought increasing interest in their structure and roles, especially in the relation to biomedicine. Thus, new computational methods dedicated to these motifs started to appear. Most of them support quadruplex analysis on the sequence level. Some touch the 3D structure and its classification defined in [Webba da Silva, 2007]. Hereby presented computational tool ElTetrado applies an approach based on the secondary structure and draws from functionality of RNApdbee 2.0 [Zok et al., 2018], our bioinformatics system for RNA secondary structure annotation. A new option in this tool to display and annotate non-canonical interactions in the secondary structure diagrams made us observe specific patterns in the visualization of RNA structures containing quadruplexes. Their observable recurrence in both RNAs and DNAs allowed for describing novel classes of tetrads and quadruplexes. We developed an algorithm to identify quadruplexes in the 3D structures and classify them according to our new nomenclature. We completed a statistical analysis of new classes' coverage by tetrads and quadruplexes included in the PDB-deposited structures. We introduced a multiline dot-bracket encoding optimized to represent the secondary structure of these motifs. Finally, we studied the relationship between our classification and the formalism defined in [Webba da Silva, 2007]. These two approaches address different properties of the structure, and – according to our discoveries – they are not totally complementary. We believe, our methodology creates a new perspective in the research focusing on quadruplex motifs and opens the new paths in their analysis.

Acknowledgements
This work was supported by grant 2016/23/B/ST6/03931 from the National Science Centre, Poland.

References

1. Webba da Silva M. (2007) Geometric formalism for DNA quadruplex folding. Chemistry, 13(35), 9738-9745.
2. Zok T., Antczak M., Zurkowski M., Popenda M., Blazewicz J., Adamiak R.W. and Szachniuk M. (2018) RNApdbee 2.0: multifunctional tool for RNA structure annotation. Nucleic Acids Research, 46(W1), W30–W35.

NeuroConstruct-based implementation of Retinal Circuitry

Miriam Elbaz[1], Rachely Butterman[1], Elishai Ezra Tsur[1]

[1]*Jerusalem College Of Technology, Jerusalem, Israel*

Neuronal modeling has proved to be indispensable to neuroscientific research as the divide between experimental and theoretical neuroscience is fading. In neuronal modeling, there is a delicate balance between bioplausibility and model tractability given rise to myriad modeling frameworks. One of the frameworks is neuroConstruct. NeuroConstruct facilitates the creation, visualization, and analysis of neural networks in 3D. Here, we extended neuroConstruct to support the generation of structured visual stimuli, to feature different synaptic dynamics, to allow for heterogeneous synapse distribution, and to enable rule-based synaptic connectivity between cell populations. We utilized this framework to demonstrate a simulation of a dense plexus of biologically realistic and morphologically detailed starburst amacrine cells. The amacrine cells were connected to a ganglion cell and stimulated with expanding and collapsing rings of light. This framework provides a powerful toolset for the investigation of the yet elusive underlying mechanisms of retinal computations such as direction selectivity, and sensitivity to moving texture, differential motion, and approaching motion.

Compute26

Managing failures in Computational Biology Workflows with PyCOMPSs

Marta Bertran[1], Jorge Ejarque[1], **Rosa M Badia[1]**

[1]*Barcelona Supercomputing Center, Barcelona, Spain*

During recent years, Computational Biology workflows are becoming bigger and more complex. These workflow are combining the use of simulators with searching and data analytics algorithms to analize a large amount of data in order to perform the desired experiments for Biologists.

One of the interesting features of the computational biology workflows is that some components of the algorithm can be tolerant to failures. Due to the nature of the algorithm, some components of the computation can fail or become blocked because of initial conditions or simulations which are not converging to a valid solution. In current workflow managers, it can make to fail or hang the whole workflow execution without getting any of the expected results. However, in some cases these computations are not critical, so the workflow could go on the rest of the computation without considering the data generated by this component in order to get a valid result. Implementing a proper failure management for this kind of failures is not currently supported for current workflow managers and implementing customized management inside the application is complicated and require important software development efforts.

PyCOMPSs [1] is a task-based parallel programming model which allow developers to easily implement parallel applications by annotating Python codes which are efficiently executed in distributed computing environments such as clusters and clouds. PyCOMPSs has been used to orchestrate complex heterogeneous workflows in different e-science fields [2][3]. In the case of life sciences, and more specific in computational biology, it is being used in the BioExcel Center of Excellence [4] to orchestrate and scale complex molecular dynamic simulations to large supercomputing infrastructures.

In this talk, we will present the PyCOMPSs programming model and how it can be used to implement computational biology workflows by mapping the different workflow computations as PyCOMPSs tasks and how the PyCOMPSs runtime automatically analyses data dependencies between tasks, detects the inherent parallelism of the application and transparently executes the tasks in the different computing resources. Apart from these features, we will also present the failure management features recently incorporated to the PyCOMPSs programming model and runtime. We will show how a developer can customize the application by defining different task properties to support failures in computational biology workflows. We will explain how to set up maximum duration time to avoid that a workflow is blocked by a non-convergent task and also how to specify hints to the runtime about what it to do when a task fails in order to complete the whole workflow execution.

[1] Tejedor, E., Becerra, Y., Alomar, G., Queralt, A., Badia, R. M., Torres, J., Cortes, T. Labarta, J. (2017). Pycompss: Parallel computational workflows in python. The International Journal of High Performance Computing Applications, 31(1), 66-82.

[2] Yildiz, O et al. (2019). Heterogeneous Hierarchical Workflow Composition. Computing in Science & Engineering.

[3] Conejero, J. et al. (2018, October). Boosting Atmospheric Dust Forecast with PyCOMPSs. In 2018 IEEE 14th International Conference on e-Science (e-Science) (pp. 464-474). IEEE.

[4] BioExcel: Centre of Excellence for Computational Biomolecular Research, https://bioexcel.eu/

Supported by

Fundació
Catalunya
La Pedrera

FUJITSU

hp

DR. ANTONI ESTEVE
FUNDACIÓ