BBCC2024

Bioinformatics and Computational Biology Conference

November 27-29, 2024

Aula Magna Facoltà di Biotecnologie

Università di Napoli Federico II

Via Tommaso de Amicis 95

Naples, Italy

PROGRAM AND ABSTRACT BOOK

BBCC2024 Conference Chairs

Dr. Angelo Facchiano – National Research Council, Institute of Food Sciences, Avellino, Italy (Chair and Coordination of the BBCC Conference Series)
Prof. Mario Guarracino – University of Cassino and Southern Lazio, Cassino, Italy.
Dr. Ilaria Granata – National Research Council, Institute for High-Performance Computing and Networking, Naples, Italy
Dr. Lucia Maddalena – National Research Council, Institute for High-Performance Computing and Networking, Naples, Italy
Prof. Anna Marabotti – University of Salerno, Italy

Scientific Committee

Dr. Claudia Angelini – National Research Council, Institute for Applied Mathematics "M. Picone", Italy
Dr. Lorraine Ayad – Brunel University London, UK
Prof. Pantelis Bagos – University of Thessaly, Greece
Prof. Maria Luisa Chiusano – University of Naples "Federico II", Naples, Italy
Prof. Angelo Ciaramella – University of Naples "Parthenope", Italy
Dr. Nunzio D'Agostino – University of Naples "Federico II", Italy
Prof. Paola Festa – University of Naples "Federico II", Italy
Prof. Vittorio Fortino – Institute of Biomedicine, School of Medicine, Faculty of Health Sciences, University of Eastern Finland, Kuopio, Finland
Prof. Christoph Friedrich – University of Applied Science and Arts Dortmund, Germany
Prof. David Gilbert – Professor (Emeritus) Computational Biology Group, Department of Computer Science, Brunel University London, UK
Dr. Giulia Guarguaglini – National Research Council, Institute of Molecular Biology and Pathology (IBPM), Rome, Italy
Dr. Bruno Hay Mele – University of Naples "Federico II", Naples, Italy
Prof. Dominik Heider – Director of the Institute for Medical Informatics, University of Münster, Germany
Prof. Maria Klapa – Institute of Chemical Engineering Sciences, Foundation for Research & Technology – Hellas (FORTH/ICE-HT)
Dr. Margherita Mutarelli – National Research Council, Institute of Applied Sciences and Intelligent Systems, Pozzuoli (NA), Italy
Prof. Romina Oliva, University of Naples "Parthenope", Italy
Dr. Vera Pancaldi, INSERM, Centre de Recherches en Cancérologie de Toulouse, France
Prof. Alessandro Pandini – Brunel University London, UK
Prof. Ahmed Rebai, Centre of Biotechnology of Sfax, University of Sfax, Tunisia
Dr. Mara Sangiovanni – University of Naples "Federico II", Naples, Italy
Prof. Antonino Staiano – University of Naples "Parthenope", Italy
Prof. Roberto Tagliaferri – University of Salerno, Italy
Prof. Ozlem Tastan Bishop – Rhodes University, South Africa
Dr. Rosa Maria Vitale – National Research Council, Institute of Biomolecular Chemistry, Pozzuoli (NA), Italy

BBCC2024 is organized with the following Partners and Sponsors:

Institute of Food Sciences of the National Research Council of Italy (ISA-CNR)

Dipartimento di Chimica e Biologia "A. Zambelli", University of Salerno, Italy

Institute for high performance computing and networking (ICAR-CNR)

BBCC is as an Affiliated Conference of ISCB since 2017

Bioinformatics Italian Society patronizes BBCC conferences since 2006

APECSA – Associazione di Promozione Educazione Comunicazione di Scienze ed Arti ETS-APS patronizes BBCC conferences since 2020

ELIXIR Italy and the ELIXIR Italy Training Group collaborate to the organization of BBCC2024 activities

Athena Consulting is partner of BBCC2024 for the administrative management of the Conference organization

computers is a sponsor of BBCC2024

NEGEDIA is a sponsor of BBCC2024

I.T.M. Informatica Telematica Meridionale s.r.l. is a sponsor of BBCC2024

BBCC2024 Main Conference Program

November 27

9:30	Opening the Registration Desk
10:30	Conference Opening – Welcome and Introduction
	Session: Structural Bioinformatics
10:50	**Invited lecture** **Marta Szachniuk** ***Frontiers in computational prediction of RNA structure***
11:40	*Francesca Ferrero* *Relevance of DNA tridimensional shape in RNA:DNA:DNA triple helix formation*
12:00	*Carmen Biancaniello* *Probing Structural Dynamics of Human Prion Protein and T183A Variant through NMR-Restrained MD Simulations*
12:20	*Nancy D'Arminio* *MDaRes: a R-driven tool for MD Data through structural alphabet approaches*
12:40	*Karol Wróblewski* *Exploring Protein Flexibility and Peptide Structure Prediction with CABS-flex*
13:00	Break – Lunch buffet
14:40	*Serena Rosignoli* *Developing Tools for Structural Bioinformatics: from Python to Bedside*
15:00	*Maria Milanesi* *Design and Validation of Broad-Spectrum Antiviral Compounds Against SARS-CoV-2*
15:20	*Simone Pirone* *Exploring Phosphomannomutase Evolution Through Structural and Sequence-Based Phylogenetics: Implications for Brain Hypoxia Response*
15:40	*Sebastian Kmiecik* *Advancing Protein-Peptide Docking: New Applications of ESMFold and CABS-dock Methods*
16:00	Coffee Break and Poster session
	Session: Databases of biological information
16:30	**Invited lecture** **Marco Beccuti** ***IT Infrastructure and Computational Services in the PNRR IR SUS-MIRRI.IT Project to support the Italian Microbial Research***
17:20	*Elisa Mauriello* *Collating marine metagenomics resources*
17:40	*Ivan Fruggiero* *An interactive genetic fingerprinting database for chestnut genotyping*
18:00	*Closing remarks of first day*

November 28

9:00	***Session: Omics and disease***
9:00	*Bruno Giovanni Galuzzi* *Identification of miRNA Biomarkers for Inflammatory Bowel Disease Using Machine Learning*
9:20	*Carmen Marino* *Metabolomic approach to investigating Nusinersen neurometabolic effects*
9:40	*Francesco Reggiani* *Data Fusion applications for cancer genomics data analysis*
9:50	*Mattia Fanelli* *Combined MERFISH and bulk-RNA seq analysis on PDAC Spheroids infected with oncolytic virus SG33*
10:00	*Francesco Massaini* *Spatial Profiling of the Tumor Microenvironment: A comparison of tools for the extraction of features predictive of therapy response*
10:10	**Invited lecture** **Enrico Glaab** ***Comprehensive blood metabolomics profiling analysis of Parkinson's disease***
11:00	Coffee Break and Poster session
	Session: Bioinformatics development and applications
11:30	*Vincenzo Bonnici* *PanDelos-plus: A parallel algorithm for computing genetic sequence homology in pangenomic analysis*
11:50	*Gregory Butler* *Feature Engineering for Protein Sequence Analysis*
12:10	*Maurizio Giordano* *Context-specific Essential Genes Identification and Prediction by Learning Multi-Omics and Network Data*
12:30	*Rodolfo Tolloi* *NaStrO: an ultra-rapid, open-source computing pipeline for Nanopore data*
12:50	*Ludovica Celli* *scVAR: a tool for the integration of genomics and transcriptomics from single cell RNA-sequencing data*
13:10	Break – Lunch buffet
14:20	***Special session: Collaborative Advancements in Bioinformatics: Integrating Infrastructure and Industrial Solutions***
14:20	*Francesca De Leo* *ELIXIR Infrastructure*
14:40	*Roberta Bosotti* *The National Facility for Data Handling and Analysis at Human Technopole: supporting the Italian research community*
15:00	*Sara Riccardo* *Empowering discovery: advancing life sciences through accessible genomic innovation*
15:15	*Paolo Bianco – Marco Fiorletta* *Empowering Scientific Research: Customized HPC Solutions for Optimal Performance and Efficiency*
15:30	*Michelangelo Sofo – Giuseppe Labianca* *DietAdhoc - A decision support system for nutrition specialists*
15:45	*Laura Casalino* *OASI Biobank: Advancing Asplenia Research through Integrated Bioinformatics and Collaborative Infrastructure*
16:00	Coffee Break and Poster session

16:30	***Session: Novel and challenging methodologies and big data analysis***
16:30	**Invited lecture** **Jack Tuszyński** ***Investigations of metabolic changes in cancer cells resulting from pharmacological agents and low-intensity electromagnetic fields***
17:20	*Leili Shahriyari* *Personalized Cancer Care through Digital Twin Technology: Integrating Patient-Specific Data with Quantitative Systems Pharmacology*
17:40	*Carmine Fruggiero* *inDAGO: a user-friendly graphical interface for dual RNA-seq data analysis*
18:00	*Closing remarks of second day*
20:00	*Social dinner (upon reservation)*

November 29

9:00	***Session: Novel and challenging methodologies and big data analysis (continued)***
9:00	*Antonella Prisco* *Modeling Variations in Antibody Response Magnitude and Longevity*
9:20	*Roberta Esposito* *Metagenomic analyses identify biosynthetic gene clusters of Mediterranean sponges leading to bioactive products*
9:40	*Aleksandra Swiercz* *Quality of semi-automated de novo genome assembly*
10:00	*End of session and communications*
	Session: Statistics and Artificial Intelligence in Data Analytics
10:10	**Invited lecture** **Audronė Jakaitienė** ***Predictive Analytics in Medicine and Biology***
11:00	Coffee Break and Posters
11:30	*Krzysztof Pysz* *Deep discriminative models in the detection of amyloid signaling motifs*
11:50	*Marco Benedetto* *AI-Driven Antibiotic Resistance Prediction in Hospital and Clinic Settings*
12:10	*Alessandro Esposito* *Classification and biochemical evaluation via Raman and Surface-enhanced Raman scattering spectroscopy of breast cancer cell lines expressing different levels of HER2*
12:30	*Francesca Cuturello* *Enhancing predictions of protein stability changes induced by single mutations using MSA-based language models*
12:50	*End of session and communications*
13:00	Break –Lunch buffet

14:30	***Session: Systems biology***
14:30	*Giovanni Scala* *Multi-omics data integration methods for cancer-subtyping, drug discovery and tumor-model alignment*
14:50	*Debora Dallera* *Integrative analysis of heterogeneous high-throughput transcriptomic data for promoter selection in bacterial genomes to support microbial synthetic biology*
15:10	*Silvia Giulia Galfrè* *Machine learning and explainable AI for transcriptomic analysis in Multiple Sclerosis*
15:30	*Chiara Cimolato* *Mathematical Modeling of Phage-Mediated CRISPRi System for Inhibiting Antibiotic Resistance*
15:50	***Special session: Funding opportunities for young scientists*** ***Organized by youngBITS, young-Infolife, and ISCB RGS-Italy groups*** *Ermanno Rizzi - Italian Cystic Fibrosis Research Foundation (FFC Ricerca)* *″Funding initiatives for young researchers: the experience of the Italian Cystic Fibrosis Research Foundation (FFC Ricerca)″.*
16:00	*Daniela Guidone - Telethon Institute of Genetics and Medicine (TIGEM) and ″Gianni Mastella Research Fellowship 2024″* *Airway surface as a battleground against bacteria*
16:10	*Michele Genovese - Telethon Institute of Genetics and Medicine (TIGEM) and ″Gianni Mastella Starting Grant 2024″* *″Alternative therapeutic target to restore the mucociliary clearance in CF″*
16:50	*Round table with the session speakers and representatives of the youngBITS, young-infolife, and RGS-Italy ISCB groups*
16:50	*Announcements: best oral and poster presentation awards – Future works*
17:00	*Closing of the Conference*

Pre-Conference Training Activities

Two training activities are held at:
Consiglio Nazionale delle Ricerche (CNR), Via Pietro Castellino 111, Naples, Italy
Each activity is restricted to a maximum number of 15 participants, selected on the basis of demand.

November 25

One-day training activity
From raw matrices to differential expression/methylation patterns: a functional genomics approach to detect molecular insights

Instructors and Organizers:
Dr. Luca Ambrosino and Dr. Francesco Cecere
Institute of Genetics and Biophysics "Adriano Buzzati-Traverso"
Consiglio Nazionale delle Ricerche, Naples, Italy

November 26

One-day training activity
Software Environments, Containers, and Notebooks (for Bioinformatics and Computational Biology)

Instructor
Dr. Raoul Bonnal, IFOM - Research Computing & Data Science Manager
Helpers
Dr. Riccardo Lorenzo Rossi, Bioinformatics scientist, and Dr. Cristiano Petrini, Bioinformatics engineer - IFOM-ETS, Research Computing & Data Science

Organisers
Prof. Anna Marabotti, University of Salerno, ELIXIR-IT, Italy
Prof. Allegra Via, University of Rome “La Sapienza”, ELIXIR-IT, Italy
Dr. Angelo Facchiano, CNR-ISA, Avellino, Italy
Dr. Ilaria Granata, CNR-ICAR, Naples, Italy
Dr. Lucia Maddalena, CNR-ICAR, Naples, Italy

This training is organized by the ELIXIR-IT Training Platform

ABSTRACTS OF INVITED LECTURES

Frontiers in computational prediction of RNA structure

Marta Szachniuk

Institute of Computing Science, Poznan University of Technology, Poland

Institute of Bioorganic Chemistry PAS, Poznan, Poland

RNA molecules play essential roles in cellular processes, from gene regulation to catalysis, and understanding their three-dimensional structures is crucial for elucidating their functions and interactions. Despite significant progress in computational methods, the sequence-based prediction of RNA 3D structures remains a substantial challenge. In this presentation, I will highlight the key challenges faced by modern predictive methods, including accurate modeling of modules stabilized by non-canonical interactions, long-range interactions, RNA complexes with other molecules, and topological puzzles. I will also outline the fundamental differences between protein and RNA structure prediction, emphasizing their unique complexities. Additionally, I will discuss the benchmarking of prediction methods and the flagship experiments, such as CASP and RNA-Puzzles. Finally, I will showcase examples of tools and techniques for modeling and analyzing 3D RNA structures.

IT Infrastructure and Computational Services in the PNRR IR SUS-MIRRI.IT Project to support the Italian Microbial Research

Marco Beccuti

University of Turin, Italy

The Italian Node of the Microbial Resource Research Infrastructure (MIRRI-IT, https://www.mirri-it.it/) plays an important role in advancing microbial research excellence in Italy. Serving as a central hub, MIRRI-IT promotes the sharing of information, data, and services related to Italian microbial collections, encouraging cooperation and idea-sharing among various stakeholders. In recognition of its strategic significance, the Italian Government funded the PNRR IR SUS-MIRRI.IT project (\url{www.sus-mirri.it}) in 2022 to enhance and secure the long-term sustainability of MIRRI-IT.
As part of this initiative, the SUS-MIRRI.IT project focuses on creating an Italian Collaborative Working Environment (ItCWE) platform. This platform will act as the core for publishing and sharing the offered information, data, and services.

This presentation details the progress in developing this open-source ItCWE platform, highlighting its key features such as the hardware and software architecture, and application layers that constitute its technical backbone.

Decoding the Blood Metabolome in Parkinson's Disease: A Systems-Level Analysis

Enrico Glaab

with

Elisa Gómez de Lope (1), Rebecca Ting Jiin Loo (1), Armin Rauschenberger (1), Muhammad Ali (1), Lukas Pavelka (2,3), Tainá M Marques (3), Clarissa P C Gomes (4), Rejko Krüger (2,3,4), Enrico Glaab (1) on behalf of the NCER-PD Consortium

(1) Biomedical Data Science, Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, Esch-sur-Alzette, Luxembourg (2) Parkinson's Research Clinic, Centre Hospitalier de Luxembourg (CHL), Luxembourg, Luxembourg (3) Transversal Translational Medicine, Luxembourg Institute of Health (LIH), Strassen, Luxembourg (4) Translational Neuroscience, Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, Esch-sur-Alzette, Luxembourg

Parkinson's disease (PD) is a complex neurodegenerative disorder characterized by significant heterogeneity in its clinical presentation and underlying molecular pathology. Despite ongoing research efforts, effective disease-modifying therapies and robust early-stage biomarkers remain elusive. To address this gap, we conducted a comprehensive, cohort-wide blood plasma metabolomic profiling study within the Luxembourg Parkinson's Study, aiming to elucidate disease-associated changes at the level of systemic cellular processes and molecular networks. Our study revealed significant changes in metabolite levels and cellular pathway activities in both dopaminergic-treated and untreated PD patients compared to controls. A key finding was the detection of coordinated increases in xanthine metabolism metabolites, particularly in early-stage, untreated (de novo) patients. These changes were consistent with independent transcriptomic measurements. Through network-based integration of metabolomics and transcriptomics data, we identified hypoxanthine phosphoribosyltransferase 1 (HPRT1) as a potential key regulator of the observed metabolite changes, potentially linking them to pathological ATP loss in PD. The novelty of our approach lies in its network-based multi-omics integration and the specific consideration of early de novo patients in addition to those receiving dopaminergic medications. The results open potential new avenues for biomarker and drug target discovery in PD, with HPRT1 emerging as a promising candidate for further investigation, particularly for early disease modification.

Investigations of metabolic changes in cancer cells resulting from pharmacological agents and low-intensity electromagnetic fields

Jack Tuszynski

DIMEAS, Politecnico di Torino, Torino, Italy
and
Department of Physics and Department of Oncology, University of Alberta, Edmonton, Alberta, Canada

Electrical and electromagnetic interactions with cells are particularly poorly understood. We report a comprehensive investigation into the effects of two distinct electromagnetic modalities on multiple cancer cell lines. The first modality involves near infrared pulsating photobiomodulation (PBM) source and the second, visible light with a hyperpolarization mechanism. Our study reveals a complex interplay between light expo- sure parameters, cellular characteristics, and metabolic reprogramming of cancer cells exposed to these sources of electromagnetic waves. While cancer cells exhibit major changes in their electro-chemical properties compared to normal cells, this property is yet to be substantially exploited for therapeutic applications. The Bioptron device, which generates "hyperpolarized light" through a unique polarization pattern, exhibited distinct cellular responses compared to the Vielight NeuroPro device. Immunofluorescence analysis revealed cell line-specific morphological alterations, including cytoplasmic shrinkage, changes in actin distribution, and potential mitochondrial damage. These structural changes were more pronounced in cells exposed to the Bioptron device, particularly after 10 minutes of exposure. Metabolic assessments indicated a shift in energy production pathways following irradiation. Some experimental sets showed increased glycolytic activity with reduced mitochondrial ATP production, while others demonstrated the opposite trend. This metabolic reprogramming appeared to be influenced by both the irradiation conditions and the specific cell line. Interestingly, experiments involving the irradiation of culture medium alone suggested that the medium plays a crucial role in mediating the effects of electromagnetic waves on cells. This finding highlights the complexity of PBM mechanisms and the potential involvement of extracellular factors in cellular responses to light exposure. The study also observed that the efficacy of PBM treatment appeared to follow the Arndt-Schultz law or the hormesis principle, where low doses stimulated cellular processes while high doses exerted inhibitory effects. This biphasic response underscores the importance of optimizing treatment parameters to achieve desired therapeutic outcomes. We also briefly discuss efforts to develop cancer chemotherapy drugs aimed at achieving a reverse Warburg effect, namely upregulation of oxidative phosphorylation and downregulation of glycolysis. Two specific drug candidates: DCA and 3BP will be discussed.

Predictive Analytics in Medicine and Biology

Audrone Jakaitiene

Vilnius University

The omics field is experiencing an explosion of data, thanks to advances in high-throughput sequencing. This abundance of data provides unprecedented opportunities for predictive modelling in precision medicine. Predictive modelling, also known as predictive analytics, uses statistical, data mining, and artificial intelligence techniques that can be applied to a variety of applications. In medicine, a predictive model learns from patients‘ historical data to predict their future outcomes and determine possible treatments. Predictive modelling holds great promise for advancing personalized medicine and improving our understanding of biological systems. However, with the opportunities, we also face challenges in proper data analysis, trustworthy interpretation and effective implementation. That is why it is so important to train new interdisciplinary teams of data scientists and clinicians to harness the full potential of predictive analytics.

ABSTRACTS OF ORAL PRESENTATIONS

Relevance of DNA tridimensional shape in RNA:DNA:DNA triple helix formation

Francesca Ferrero, Marco Masera, Chiara Cicconetti, Ivan Molineris

Università degli studi di Torino, DBIOS, via Accademia Albertina 13 Torino (Italy)

Long non-coding RNAs (lncRNAs) play key roles in regulating gene expression by interacting with DNA through various mechanisms. One such mechanism is the formation of RNA:DNA:DNA triple helices (triplex), where a single strand of RNA forms non-canonical hydrogen bonds with double-stranded DNA, interacting with the major groove of the genomic helix (Maldotti 2022). Predicting these triplexes computationally has gained significant attention in recent years, but consensus on optimal prediction methods remains elusive (Cicconetti 2023). This study investigates the influence of DNA 3D shape features — helix twist (HelT), minor groove width (MGW), propeller twist (ProT), and roll (Roll) — on the formation of triplexes and their potential to improve prediction models. To explore this, we collected experimentally validated triplex targets and estimated the values for the four shape features at binding sites and control regions. We employed nested logistic regression models to assess whether incorporating these DNA shape features, alongside traditional sequence-based data, enhances prediction accuracy. Our models demonstrated that adding 3D shape parameters increased the area under the curve (AUC) of the predictions by 17.5%, representing a significant improvement in predictive performance, even after controlling for chromatin openness, a factor known to enrich triplex sites. Further analysis revealed that the most significant contributions came from the HelT and ProT features, which showed the greatest difference between triplex-forming and non-forming regions. The MGW and Roll parameters, while still contributing to the model, had a smaller but noticeable impact. Importantly, the 3D DNA shape features improved triplex prediction across various lncRNAs, demonstrating that these features offer independent and complementary information to traditional sequence-based predictions. These findings indicate that the tridimensional structure of DNA plays a critical role in determining RNA:DNA:DNA triplex formation. By incorporating DNA shape features into triplex prediction algorithms, such as 3plex (Cicconetti 2023), we can achieve more accurate and reliable results, potentially opening up new avenues for understanding the regulatory roles of lncRNAs in gene expression and their broader implications in health and disease.

REFERENCES

Maldotti M, Lauria A, Anselmi F, Molineris I, Tamburrini A, Meng G, Polignano IL, Scrivano MG, Campestre F, Simon LM, Rapelli S, Morandi E, Incarnato D, Oliviero S. The acetyltransferase p300 is recruited in trans to multiple enhancer sites by lncSmad7. Nucleic Acids Res. 2022. doi: 10.1093/nar/gkac083.

Cicconetti C, Lauria A, Proserpio V, Masera M, Tamburrini A, Maldotti M, Oliviero S, Molineris I. 3plex enables deep computational investigation of triplex forming lncRNAs. Comput Struct Biotechnol J. 2023. doi: 10.1016/j.csbj.2023.05.016.

Probing Structural Dynamics of Human Prion Protein and T183A Variant through NMR-Restrained MD Simulations

Carmen Biancaniello, Michail D. Vrettas, Alessandro Emendato, Alfonso De Simone

University of Naples Federico II, Italy

Prion protein is a monomeric glycoprotein primarily expressed in the nervous system. Its normal cellular form (PrPC) consists of a flexible and disordered N-terminal tail and a globular C-terminal domain composed of three helices (H1-H3) and a small beta-sheet. Unfortunately, PrP is known to misfold from the PrPC form into the pathological scrapie form (PrPSc), which aggregates into insoluble amyloid fibrils causing transmissible spongiform encephalopathies (TSEs), a group of fatal neurodegenerative disorders. In humans, TSEs can be inherited through mutations in the PRNP gene, with the T183A mutation associated with a familial form of TSE, Creutzfeldt-Jakob disease, and reported to have a strong destabilizing effect by disrupting a key hydrogen bond within the C-terminal domain. Despite extensive experimental and theoretical investigations, the mechanism underlying the misfolding and aggregation of PrPC remains elusive. The main limitation lies in the proper characterization of monomeric, partially unfolded species formed during the transition from PrPC to PrPSc, which is difficult to achieve with experimental approaches due to their transient and heterogeneous nature. Computational techniques offer a valuable alternative to address this challenge, but they depend on the limited accuracy of empirical force fields, which can hinder the identification of these metastable states. In this study, we performed replica averaged NMR-restrained MD simulations of the C-terminal globular domain of both human PrP and the T183A variant using our previously developed method, NapShift. This method employs artificial neural networks to predict chemical shifts (CS) from structural data and enables derivatives to apply experimental CS as restraints, improving the quality of biomolecular simulations. The analysis of the structural ensembles revealed the native conformation as the most populated state for the wild-type protein. In contrast, three additional, less-populated states were observed for the mutant, suggesting a higher propensity to misfold. The alternative conformational clusters featured a stable subdomain comprising helices H2 and H3, consistent with experimental studies, and a progressive destabilization of the regions including H1 and the beta-sheet. NMR relaxation measurements aligned with our computational results validating our method’s ability to detect intermediate states that may act as key drivers of the misfolding process. These findings offer new insights into the dynamical behavior of PrP, which can inform the design of more effective therapeutic strategies for prion diseases.

MDaRes: a R-driven tool for MD Data through structural alphabet approaches

Nancy D’Arminio (1), Anna Marabotti (1)$, Alessandro Pandini (2)$

1: Department of Chemistry and Biology “A. Zambelli”, University of Salerno, Fisciano (SA), Italy
2: Department of Computer Science, Brunel University London, Uxbridge UB8 3PH, United Kingdom.
$: co-corresponding authors

Analyzing molecular dynamics (MD) data is key to understanding the dynamic nature of proteins, predicting their interactions, guiding drug discovery efforts, and unraveling the mechanisms underlying their functions. MDaRes, a tool developed in R, enables the detailed analysis of MD data through the use of structural alphabets. The software offers three core functionalities. First, it provides in-depth analysis of local structural changes within proteins, which helps clarify the dynamic behavior of specific regions. Second, it identifies and assesses allosteric communication pathways, revealing how structural changes in one region can influence distant parts of the protein. Third, it pinpoints key regions crucial for the overall movement and function of the protein, assisting in the design of specific protein variants for further research. MDaRes enables the analysis of protein function under different conditions, such as ligand binding or mutations, giving users a comprehensive molecular perspective. It is built on the M32K25 structural alphabet [PMID: 20170534], which consists of 25 canonical fragment states, each representing a group of four residues focused on Calpha atoms. By encoding protein conformations through matching fragments based on Root Mean Square Deviation (RMSD), the software compresses a protein with n residues into a structural string of length n-3. Residue annotations within MDaRes are enriched with data from the Atlantis database (https://atlantis.bioinfolab.sns.it/), which consolidates information from various authoritative sources like PDBe, Pfam, InterPro, UniProt, IntAct, PDB, and AlphaFold. This ensures that the structural and functional insights generated by MDaRes are rooted in current, reliable biological information. The development and validation of MDaRes involved molecular dynamics simulations of 600 nanoseconds (with three replicas) at 310K on two case systems: the wildtype galactose-1-phosphate uridyltransferase enzyme (wtGALT) and its variant p.Gln188Arg, using the GROMACS platform. Through an analysis of fragment frequency, Shannon Entropy, and mutual information between fragment pairs, the output from MDaRes was confirmed to align with existing scientific knowledge. MDaRes is an open-source, user-friendly tool designed to handle the extensive datasets typical of MD simulations. It incorporates robust statistical tools for comprehensive analysis, making it highly valuable for studying protein dynamics and aiding in the design of protein variants. Its flexibility and accessibility through R ensure it is an effective and intuitive resource for researchers exploring the complex dynamics of biological systems.

Exploring Protein Flexibility and Peptide Structure Prediction with CABS-flex

Karol Wróblewski, Sebastian Kmiecik

Biological and Chemical Research Centre, Faculty of Chemistry, University of Warsaw, Warsaw, Poland

Revolution in deep learning and methods like AlphaFold have significantly impacted the field of protein and peptide structure prediction [1]. Because training these models relies on large datasets of static protein structures, the subsequent methods typically provide only a single, static conformation for each query sequence. However, capturing the dynamic nature of protein structures requires tools that can also model flexibility. Here, we introduce new extensions to CABS-flex, an established tool for fast simulations of protein and peptide flexibility [2], [3], [4]. First, we present a new feature of CABS-flex: the ability to predict structures of both linear and cyclic peptides. By leveraging diverse datasets of short peptides, we evaluate CABS-flex alongside various state-of-the-art methods [5]. While AlphaFold is highly effective at predicting static structures of larger proteins, our results show that CABS-flex proves especially effective for peptides, where increased flexibility is a critical factor. Second, we extend CABS-flex with a novel approach that incorporates AlphaFold's pLDDT confidence scores and secondary structure into our flexibility prediction pipeline. We validate this approach by benchmarking it against molecular dynamics (MD) simulations of approximately 1,400 proteins. We demonstrate that addition of this new information significantly improves accuracy of flexibility modeling.

References:

[1] J. Jumper et al., "Highly accurate protein structure prediction with AlphaFold," Nature, vol. 596, no. 7873, pp. 583–589, Aug. 2021, doi: 10.1038/s41586-021-03819-2.

[2] A. Kuriata et al., "CABS-flex 2.0: a web server for fast simulations of flexibility of protein structures," Nucleic Acids Res., vol. 46, no. W1, pp. W338–W343, Jul. 2018, doi: 10.1093/nar/gky356.

[3] M. Kurcinski, T. Oleniecki, M. P. Ciemny, A. Kuriata, A. Kolinski, and S. Kmiecik, "CABS-flex standalone: a simulation environment for fast modeling of protein flexibility," Bioinformatics, vol. 35, no. 4, pp. 694–695, Feb. 2019, doi: 10.1093/bioinformatics/bty685.

[4] C. Nithin et al., "Exploring protein functions from structural flexibility using CABS ?flex modeling," Protein Sci., vol. 33, no. 9, p. e5090, Sep. 2024, doi: 10.1002/pro.5090.

[5] A. Badaczewska-Dawid, K. Wróblewski, M. Kurcinski, and S. Kmiecik, "Structure prediction of linear and cyclic peptides using CABS-flex," Brief. Bioinform., vol. 25, no. 2, p. bbae003, Mar. 2024, doi: 10.1093/bib/bbae003.

Developing Tools for Structural Bioinformatics: from Python to Bedside

Serena Rosignoli (1), Dalila Boi (2), Elisabetta Rubini (2), Italia Anna Asteriti (2), Giulia Guarguaglini (2) and Alessandro Paiardini (1)

(1) Department of Biochemical Sciences, Sapienza University of Rome, P. le Aldo Moro 5, 00185 Rome, Italy
(2) Institute of Molecular Biology and Pathology, National Research Council of Italy, c/o Sapienza University of Rome, Via degli Apuli 4, 00185 Rome, Italy.

Software engineering in protein structural biology has evolved from solving individual problems to embracing the FAIR (Findable, Accessible, Interoperable, Reusable) principles [1, 2], ensuring software is both scalable and adaptable to the broader bioinformatics landscape. In this context, we are dedicated to the development of tools to assist structural bioinformatics with user-friendly environments. PyMod [3] and DockingPie [4], both PyMOL plugins, are designed to facilitate protein sequence and structure analysis and protein-ligand docking, respectively. By building on PyMOL's powerful capabilities, these tools create a streamlined workflow, simplifying everyday tasks in protein structural analysis. Our research exemplifies the application of these tools, enabling us to target key oncogenic drivers in neuroblastoma through peptide design and small-molecule inhibitors [5]. With AlPaCas [6], a web-based platform to identify and optimize Cas proteins adept at allele-specific targeting, we have employed protein engineering to develop Cas proteins for allele-specific targeting of pathogenic mutations in epidermolysis bullosa. This platform makes genome-wide analysis and protein engineering accessible to different levels of expertise. The rapid growth of biological (e.g. structural) data has driven advancements in AI, reinforcing the need for flexible, interoperable tools capable of tackling increasingly complex challenges [7]. While AI tools are significantly influencing standard protocols in the field, a unified environment serves as a crucial bridge in cases where these newer tools still fall short. This ensures continuity and adaptability, allowing researchers to seamlessly integrate both traditional and AI-driven methods.

References

[1] Gauthier J, et al. (2019) Briefings in Bioinformatics
[2] Barker, M. et al. (2022) Sci Data
[3] Janson, G. and Paiardini, A. (2021) Bioinformatics
[4] Rosignoli, S. and Paiardini, A. (2022) Bioinformatics
[5] Boi, D. et al. (2021) Int J Mol Sci
[6] Rosignoli, S. et al. (2024) Nucleic Acids Research
[7] Rosignoli, S. et al. (2024) FEBS Open Bio.

Design and Validation of Broad-Spectrum Antiviral Compounds Against SARS-CoV-2

Maria Milanesi (1,2), Chiara Urbinati (2), Liv Zimmermann (3), Pasqua Oreste (4), Francesca Caccuri (5), Petr Chlanda (3), Rebecca C. Wade (7,8,9), Giulia Paiardi (7,8), and Marco Rusnati (2,10)

(1) Institute for Biomedical Technologies, National Research Council (ITB-CNR), 20054 Segrate (MI), Italy;
(2) Macromolecular Interaction Analysis Unit, Section of Experimental Oncology and Immunology, Department of Molecular and Translational Medicine, 25123 Brescia, Italy
(3) Schaller Research Group, Department of Infectious Diseases-Virology, Heidelberg University, 69120, Heidelberg
(4) Glycores 2000 S.r.l., Milan, Italy
(5) Section of Microbiology, Department of Molecular and Translational Medicine, University of Brescia, Brescia, Italy
(6) Infectious Diseases Imaging Platform (IDIP), Department of Infectious Diseases, Heidelberg University, Heidelberg, Germany
(7) Molecular and Cellular Modeling Group, Heidelberg Institute for Theoretical Studies (HITS), 69118 Heidelberg, Germany
(8) Zentrum für Molekulare Biologie (ZMBH), Heidelberg University, 69120 Heidelberg, Germany
(9) Interdisciplinary Center for Scientific Computing (IWR), Heidelberg University, 69120 Heidelberg, Germany
(10) Consorzio Interuniversitario Biotecnologie (CIB), Unit of Brescia, 25123 Brescia, Italy

The impact of SARS-CoV-2 has underscored the significant threat posed by viral infections, with the rapid emergence of variants highlighting the urgent need for the development of broad-spectrum antivirals. In this context, the primary molecular target of SARS-CoV-2 is its spike protein trimer, located on the viral envelope, which mediates the virus's attachment and entry into host cells by interacting with the angiotensin-converting enzyme 2 (ACE2) receptor and heparan-sulfated proteoglycans (HSPGs) on the cell surface. Recent studies have revealed the crucial role of HSPGs in the viral infection process, indicating that their antagonists may hold promise as broad-acting anti- SARS-CoV-2 candidates. In fact, treatment with heparin, a structural analogue of HS, has been shown to significantly reduce SARS-CoV-2 infection rates. However, since heparin is primarily utilized for its anticoagulant properties, it can lead to significant and unwanted side effects, limiting its application as an antiviral therapy. To address this issue, we aimed at design and validate a library of heparin analogues that retain antiviral activity but are devoid of anticoagulant side effects. Our multidisciplinary approach combines (i) biochemical assays, such as Surface Plasmon Resonance (SPR) and microscale thermophoresis (MST), to evaluate the binding affinity of the compounds for the spike protein and their inhibitory capability regarding HS and ACE2 binding; (ii) extensive molecular dynamics (MD) simulations of the spike protein in complex with the designed HS analogues, highlighting the mechanisms of binding and inhibition; and (iii) cell-based experiments against various SARS-CoV-2 variants to confirm their antiviral potential and broad-spectrum activity. Our results indicate that two of the designed heparin derivatives exhibit greater antiviral activity than heparin, due to their enhanced ability to inhibit the direct binding of the spike protein to HS and ACE2. These findings support the therapeutic potential of these heparin analogues as antiviral agents with improved efficacy and reduced side effects compared to heparin.

Acknowledgements: This activity has been supported by the funding of the European Union (NextGenerationEU), Italian NRRP project code IR0000031 - Strengthening BBMRI.it - CUP B53C22001820006.

Exploring Phosphomannomutase Evolution Through Structural and Sequence-Based Phylogenetics: Implications for Brain Hypoxia Response.

Simone Pirone (1), Maria Monticelli (1,2), Giuseppina Andreotti (2), Maria Vittoria Cubellis (1), Bruno Hay Mele (1)

(1) Biology Dept. ; Università degli Studi di Napoli Federico II; Complesso Universitario MSA, Ed. 7; Via Cinthia, 26 80126 Naples (NA)
(2) Institute of Biomolecular Chemistry ICB, CNR, Via Campi Flegrei 34, 80078, Pozzuoli, Italy

Phosphomannomutase-1 (PMM1) and phosphomannomutase-2 (PMM2) are two paralogous enzymes that share mutase activity in vitro, but show an interesting functional divergence in vivo. In humans, PMM2 is activated by glucose-1,6-bisphosphate (Glc-1,6-P2), and acts as a phosphomannomutase, converting mannose-6-phosphate (Man-6-P) to mannose-1-phosphate (Man-1-P). Its deficiency underlies a rare genetic disease, PMM2-CDG (congenital disorder of glycosylation). In addition to mutase activity, PMM1 has phosphatase activity on Glc-1,6-P2. This is particularly relevant in the brain, where, in case of hypoxia the phosphatase activity is activated by inosine monophosphate (IMP). Given the evidence that PMM1 does not compensate for the lack of PMM2 in PMM2-CDG patients, the study of their functional divergence is of particular interest. Unraveling the evolution of these two paralogues could shed light on their functional divergence and biological significance in vivo. In this study, we performed phylogenetic analysis on sequences of metazoan PMM orthologues. The resulting phylogenetic tree shows the division of PMMs into three clusters; however, the position of osteichthyan taxa PMM2 is unexpected. Furthermore, residues crucial to IMP binding are conserved in PMM1 but not in PMM2, and mammals PMM1 present an interestingly conserved N-term. Additionally, we compared this sequence-based phylogeny with one produced using THESEUS to assess whether structural signals are consistent with sequence-based evolutionary relationships. The pivotal role that molecular oxygen has in metazoans, especially in the brain of higher organisms, and the peculiar hypoxia-induced bisphosphatase activity of PMM1 lead to the hypothesis that PMM1 might be under selective pressure in the brain. This pressure would result in genetic convergence in other organisms, as evidenced in osteichthyes. Comparing the average degree of sequence similarity between orthologs in the two lineages, it appears that PMM1 have a slightly higher average degree of identity to each other than PMM2. This evidence, together with the high mutation rate of PMM1, supports our hypothesis.

Advancing Protein-Peptide Docking: New Applications of ESMFold and CABS-dock Methods

Sebastian Kmiecik

Biological and Chemical Research Centre, Faculty of Chemistry, University of Warsaw

Accurate peptide docking remains a key challenge in the development of peptide therapeutics. Over the past few years, we have focused on protein-peptide docking and developed the well-established CABS-dock tool. Recently, rapid advancements in deep learning have introduced new opportunities in this field. In this presentation, I will discuss our recent assessment of the ESMFold language model, originally designed for protein structure prediction, to evaluate its effectiveness in protein-peptide docking. We explored multiple docking strategies, including poly-glycine linkers and sampling-enhancing modifications. Our findings show that the number of acceptable-quality models produced by ESMFold is comparable to traditional methods and, in some cases, surpasses AlphaFold. The combination of result quality and computational efficiency makes ESMFold a valuable tool for high-throughput peptide design. Additionally, I will introduce a novel extension of the CABS-dock method designed for cyclic peptide docking, which further expands its capabilities in therapeutic peptide development.

Collating marine metagenomics resources

Italia Elisa Mauriello (1) Maria Chiara Langella (1) Marco Miralto (2) Edoardo Pasolli (1) Maria Luisa Chiusano (1)

(1) Department of Agricultural Science - Università degli Studi di Napoli Federico II - Reggia di Portici - Via Università, 100 - 80055 - Portici (NA)
(2) Department of Research Infrastructures for Marine Biological Resources (RIMAR) - Stazione Zoologica Anton Dohrn - Via Francesco Caracciolo - Villa comunale - 80122 – Naples

Marine metagenomics has rapidly expanded due to improved sampling and sequencing technologies that and the emergence of wide interest for emerging applications ranging from biodiversity assessment to ecological monitoring and bioprospecting. Indeed, microorganisms dominate marine habitats in terms of total biomass and metabolic activity playing important roles in various biogeochemical and environmental processes. Because of the opportunities to understand the complexity of marine microbial ecosystems, marine metagenomics has gained traction through several worldwide campaigns. Moving from the initial challenge proposed by Craig Venter in 2004 (Venter et al., 2004), aimed to the whole-genome sequencing applied to seawater microbial samples from the Sargasso Sea, and the following Sorcerer II Global Ocean Sampling (GOS) Expedition (Rusch et al. 2007), more ambitious efforts followed, among these the Tara Oceans Expedition (2009–2013) (Karsenti et al., 2011) and the Malaspina Expedition (2010-2011) (Duarte, 2015). Despite the significant investments, extensive data collection, and impressive project-specific results, a major challenge remains in organizing and managing the whole amount of data by a global view. The current digital organization of data resources—while raw sequences are centralized in global reference collections such as NCBI (Benson et al. 1990), EMBL-EBI (Emmert et al., 1994), and DDBJ (Tateno et al., 1997)—still falls short of providing immediate access to reusable insights from the various initiatives. As a result, data from these efforts remain scattered across project-dependent repositories, creating obstacles for both inexperienced and expert users attempting to access the value added still hidden information. These efforts, accompanied by more recently launched projects aiming at long-term omics observations of marine diversity, (like the EMO BON project by the European Marine Biological Resource Centre-EMBRC 2021) (Santi et al., 2021) contribute significant materials for expanding the comprehension of the magnitude of the oceans' microbial content. To facilitate the tracking of metagenomics resources from key campaigns, we conducted a review of the main publicly available metagenomics reference databases. Our focus was specifically on major marine metagenomics projects, to track raw data resources and associated processed results, to favour the access of the interested community. To maximize the potential of marine metagenomic data, our effort paves the way to the next challenge aimed to implement adequate platforms and associated technologies for a one stop access favouring the exploitation of the whole amount of data by integrative approaches.

An interactive genetic fingerprinting database for chestnut genotyping

Ivan Fruggiero (1), Alessandro Maisto (1), Sara Passaro (2), Domenico Gentile (2), Angelina Nunziata (2), Nunzio D’Agostino (1)

(1) Department of Agricultural Sciences, University of Naples Federico II, piazza Carlo di Borbone 1, 80055, Portici, Napoli, Italy (2) Council for Agricultural Research and Economics (CREA), Research Center for Olive, Fruits, and Citrus Crops, 81100, Caserta, Italy

Background: The European chestnut (Castanea sativa Mill., Fagaceae) holds considerable ecological and economic significance, particularly in countries such as Italy, Greece, Spain and Turkey. This species plays a crucial role in rural economies and ecosystems. Accurate varietal recognition is essential for managing chestnut groves, influencing everything from planting to marketing. Traditional morpho-physiological methods face limitations, particularly in distinguishing young, dormant, or scion trees, and require extensive expertise, which is hard to transfer and often costly. These challenges can determine investment and impact varietal choices. Recent advancements, such as Single Nucleotide Polymorphisms (SNPs) markers identified through KASP technology, have greatly improved cultivar identification. The availability of KASP-based genotyping data has significantly accelerated the development of a database of genetic fingerprints to assist farmers and researchers in accurately identifying and distinguishing chestnut varieties. Methods: To develop the database, we first assessed user needs to determine required data and query functionalities. We then created an Entity-Relationship (ER) diagram for the database structure and implemented it using MySQL. The interactive platform was built with PHP, HTML, CSS, and JavaScript, and integrated bioinformatics tools like BLAST, MAFFT, and BioPython for enhanced data analysis and user experience in genetic fingerprinting and chestnut genotyping. Results: The database includes genotypic and phenotypic data for over 150 chestnut genotypes, featuring 38 SNPs with details on primers, sequences, positions, and variations. Users can query the database in three modes: Discriminating accessions by SNPs and/or morphological traits: Users can select SNPs and/or phenotypic traits to retrieve accessions that match the criteria, with results highlighting SNPs with the highest PIC values. This mode helps accurately distinguish varieties based on genetic and morphological data. Identifying most discriminative SNPs: This tool helps identify the SNPs that are most effective in distinguishing between recorded accessions. By selecting an accession, the platform provides a list of the minimum number of SNPs and associated alleles that are most effective for its discrimination and characterization. This feature is especially valuable for developing specific KASP protocols. Searching user-defined genotype strings: Users can input genotype strings (concatenated SNP loci ordered by chromosome and position) to calculate the genetic distance between new and previously genotyped accessions. This function aids in the classification of new varieties by comparing their genetic profiles with existing data. Conclusions: This platform is the first interactive chestnut database for cultivar discrimination using KASP genotyping data, providing tailored feedback and a dynamic tool for cultivar identification and differentiation.

Identification of miRNA Biomarkers for Inflammatory Bowel Disease Using Machine Learning

Bruno G. Galuzzi, Fabio Pirovano, Flaminia Tani, Gloria Bertoli

(1) Istituto di Bioimmagini e Sistemi Biologici Complessi
(2) National Biodiversity Future Center

The integration between the fields of bioinformatics, Machine Learning (ML), and molecular biology opens the possibility to search for biomarkers for diagnosis, prognosis and evaluation of the effectiveness of treatment in several diseases. In this study, we consider Inflammatory bowel disease (IBD), which is a chronic and progressive immune-mediated inflammatory condition, characterized by inflammation of the gastrointestinal tract. This disease is often hard to manage in clinical practice due to the high variability of its clinical manifestations and their multifactorial etiology. However, several studies examining the association between IBD and epigenetic molecules (e.g., miRNAs and genes) have linked the dysregulation of these molecules to key signaling pathways involved in the pathogenesis of IBD. This, associated with the large volume of biological data publicly available nowadays (e.g. microarray and next-generation sequencing experiments), makes it possible to explore various ML techniques, with the aim of identifying potential biomarkers for disease prediction and/or patient stratification. In this study, we want to explore the possibility of using supervised ML models to identify those biomarkers able both to early predict IBD and those able to diagnose IBD. To train and test different ML models, we retrieved several microarray datasets from the Gene Expression Omnibus database, containing miRNA expression profiles of individuals with Ulcerative Colitis (UC), Crohn's Disease (CD)—the two main clinical forms of IBD—and healthy individuals. A subset of miRNAs was extracted from an initial statistical analysis to select the most differentially expressed miRNAs between disease and healthy patient Subsequently, we will apply several classification models, among which Logistic Regression, Support Vector Machines, Random Forest and Extreme Gradient Boosting, ranking them based on the performances on the test set, using the F1-score. Finally, we will provide insights into the features (i.e., microRNAs) that most strongly affect patient stratification using SHAP analysis. Additionally, we rationalized them based on previous studies and the associated biological mechanism/pathways using Enrichment analysis, focusing on those associated with inflammation and oxidative stress.

Metabolomic approach to investigating Nusinersen neurometabolic effects

Carmen Marino (1), Francesco Errico (2,3), Valentina Bassareo (4), Valeria Valsecchi (5), Tommaso Nuzzo (3,6), Paola Brancaccio (5) , Giusy Laudati (5) , Antonella Casamassa (7), Manuela Grimaldi(1), Adele D'Amico (8), Manolo Carta (4), Enrico Bertini (8), Giuseppe Pignataro (5) , Anna Maria D'Ursi (1), Alessandro Usiello (3,6).

1) Department of Pharmacy, University of Salerno, 84084 Fisciano, Salerno, Italy
2) Department of Agricultural Sciences, University of Naples "Federico II", 80055 Portici, Italy
3) Laboratory of Translational Neuroscience, Ceinge Biotecnologie Avanzate, 80145 Naples, Italy
4) Department of Biomedical Sciences, University of Cagliari, 09042 Monserrato, Italy
5) Division of Pharmacology, Department of Neuroscience, Reproductive and Dentistry Sciences, School of Medicine, University of Naples "Federico II", 80131 Naples, Italy
6) Department of Environmental, Biological and Pharmaceutical Science and Technologies, Università degli Studi della Campania "Luigi Vanvitelli", 81100 Caserta, Italy
7) IRCCS Synlab SDN, 80143 Naples, Italy
8) Unit of Neuromuscular and Neurodegenerative Disorders, Bambino Gesù Children's Hospital IRCCS, 00163 Rome, Italy.

Spinal muscular atrophy (SMA) is a neuromuscular degenerative disease caused by homozygous deletions or mutations in the survival motor neuron 1 (SMN1) gene. Although SMA is a prototypical motor neuron disorder, findings in animal models and patients also indicate multiorgan and metabolic abnormalities. Nusinersen is the first therapeutic approved for SMA. It is an antisense oligonucleotide to be administrated by intrathecal injection, that promotes SMN protein induction. Despite the efforts to identify the biomarkers of SMA progression and therapeutic efficacy, little is known about how the disease and the treatments condition the metabolomic profile. Considering this evidence, we carried out a metabolomic study to determine the neurometabolic effects of the treatment with Nusinersen in the cerebrospinal fluid (CSF), the biofluid in which Nusinersen is administered, to guide its therapeutic effects through induction of SMN. We used Nuclear Magnetic Resonance spectroscopy (NMR) to delineate the neurometabolic signature of SMA patients' CSF according to the disease severity before and after treatment with Nusinersen. Our findings have shown that Nusinersen induces profound but distinct metabolomic changes in SMA patients of varying severity of the disease, impacting glucose metabolism for type 1 and ketone body metabolism for type 2. Additionally, we have highlighted a common dysmetabolism of aromatic amino acids (AAAs) such as phenylalanine and tyrosine. Considering the involvement of AAAs in protein synthesis, liver gluconeogenesis, ketogenesis, hormone and catecholamine synthesis, we decided to perform an untargeted metabolomic analysis to investigate whether SMN protein deficiency disrupts the liver and brain metabolomic profile in the SMN Knock-out mouse model (SMN?7) at different stages of symptomatology. Next, high-performance liquid chromatography (HPLC), quantitative RT-PCR (qrt-PCR), western blotting (WB), and immunohistochemistry (IHC), were performed to explore the consequences of SMN deficiency in influencing dopamine (DA), norepinephrine (NE) and serotonin (5-HT) metabolism of SMN?7 mice. Finally, was evaluated whether the overregulation of SMN brought by the administration of Nusinersen modulates the levels of monoamine neurotransmitter in the CSF of patients SMA1, SMA2 and SMA3 at doses of load and maintenance of therapy compared to basal. These findings reveal disease severity-specific neurometabolic signatures of Nusinersen treatment, suggesting this CNS-directed therapy's selective modulation of peripheral organ metabolism in severe SMA patients.

Data Fusion applications for cancer genomics data analysis

Francesco Reggiani (1), Max Pfeffer (2), Ulrich Pfeffer (1,3), Adriana Amaro (1)

(1) Laboratory of Gene Expression Regulation, IRCCS Ospedale Policlinico San Martino, 16132 Genova, Italy
(2) Institute of Numerical and Applied Mathematics, University of Göttingen, 37083 Göttingen, Germany
(3) corresponding author: ulrich.pfeffer@hsanmartino.it

Background: Data fusion techniques (DF) are mathematical methods applied to merge multiple genomic domain data, such as RNA-seq or methylation array dataset of tumor samples, to detect clinically relevant clusters or features that are not evident at the single domain level. In our previous works we applied DF techniques, notably Joint Singular Value Decomposition (jSVD) to Uveal and Skin Cutaneous Melanoma TCGA dataset (SKCM). jSVD analysis of SKCM RNAseq and methylation data have defined one cluster of patients characterized by increased survival and expression of immune system genes. Notably, enhanced expression of immune related signatures in cutaneous melanoma is associated with improved immune checkpoint inhibitors response and survival. Multiple data integration methods have been developed so far, most of them made available inside the Movics R package, a valuable resource to compare clusters obtained with different data fusion methods and extract related gene signatures. Methods: We integrated multi genomic SKCM data clustering from multiple data integration techniques as jSVD and Movics. Gene signatures were extracted by single methods or considering a consensus cluster defined by multiple methods: samples shared in all clusters with highest survival rates of applied DF methods were compared to the remaining and used to define a new survival signature. Genes prioritized with different data fusion methods were then compared on different datasets, to assess if detected signatures could be validated in terms of survival and response to anti PD-1 and epigenetic drugs therapy in different datasets. Results: Application of multiple data integration techniques showed the presence of overlap between clusters predicted with different methods. Gene signatures were able to predict response to therapy from RNAseq data with a performance comparable or superior to previously published immuno-signatures. Conclusions: Data fusion can be used for feature selection and identification of clusters of patients with relevant clinical features that cannot be clearly identified at the single domain level. Application of multiple data integration techniques could lead to more precise cluster definition and gene signatures extraction. Defined features can be used to classify tumors from other datasets including those with limited sample numbers or only single domain information. However, even DF captures only a part of the complex determinants of response to immunotherapy.

Acknowledgments: This research was funded by FONDAZIONE AIRC under the 5 per Mille 2018-ID.21073 program-P.I. Maio Michele to U.P.; and by the Italian Ministry of Health, Ricerca Corrente 2022 to U.P. and the Italian Ministry of Health 5X1000 2018/19 to A.A. Additionally, M.P. was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)—Projektnummer 448293816.

Combined MERFISH and bulk-RNA seq analysis on PDAC Spheroids infected with oncolytic virus SG33

Mattia Fanelli, Agathe Redouthe, Julie Bordenave, Pierre Cordelier and Vera Pancaldi

Università degli studi di Firenze, Research Center In Cancer De Toulouse, Research Center In Cancer De Toulouse, Research Center In Cancer De Toulouse.

Introduction Pancreatic cancer is one of the most aggressive malignancies, with a generally unfavorable prognosis. The associated mortality is high, and immunological therapies have been largely ineffective (Siegel et al., 2020). An innovative treatment is the use of oncolytic viruses, these can target cancer cells, inducing their death (Thorne & Contag, 2020). At present, their full therapeutic potential is hindered by a lack of knowledge. One of the most promising OVs is SG33, derived from the Myxoma virus, engineered to include the ANCHOR imaging system, which allows tracking of the virus via GFP (Cordelier et al., 2021).This approach alone does not provide sufficient understanding of cell behavior during infection. To gain further insight, we require methods that allow both spatial and transcriptional information to be gathered simultaneously. Methods We therefore employed MERFISH, a multiplexed single-molecule imaging technology for spatially resolved transcriptomics to characterise SG33 infections. We considered spheroids generated by pancreatic cancer cells infected by SG33 that were analyzed with a panel of 244 metabolic genes at several time points during the infection (mock, 24 and 48 hours post-infection). We also performed bulk RNAseq transcriptomics on 2D cultures infected with SG33 at four different time points (mock, 6, 18 and 24 hours post-infection). Results MERFISH produced estimates of transcript levels in each cell. After pre-processing, several cellular phenotypes could be identified using the Leiden hierarchical clustering method. It was however difficult to relate the isolated clusters to the presence of the virus with a high degree of confidence, mainly because of the small number of genes and samples available in our dataset. For this reason, we chose to incorporate bulk RNA-seq data to see if we could integrate the results from both techniques and gain a better understanding of the virus's effects. First, we analyzed the data using Independent Component Analysis, which allowed us to de-couple several signals hidden in the bulk transcriptomics to study underlying infection mechanisms in the form of gene lists (components) associated to specific time points or to SG33 treated vs control. We repeated the ICA analysis only focusing on the subset of genes present in the MERFISH data. Surprisingly, this approach gave us the possibility to select genes from the MERFISH panel affected by the virus.Motivated by these findings, we re-analyzed the MERFISH data, focusing on the genes we selected. We demonstrated that the presence of the virus was actually caught by the genes that we selected in the bulk analysis. These results highlight the potential of combining single-cell techniques with more traditional RNA sequencing approaches to achieve a more detailed understanding of complex biological systems.

Spatial Profiling of the Tumor Microenvironment: A comparison of tools for the extraction of features predictive of therapy response

Francesco Massaini (1), Abdelmounim Essabbar (2), Alexis Coullomb (2) and Vera Pancaldi (2)

(1) University of Florence, Italy; (2) Toulouse Cancer Research Center, Toulouse, France

The tumor microenvironment (TME), a complex mix of cells surrounding cancer cells, significantly impacts tumor progression and treatment responses. We present initial steps toward multi-omics analysis of the TME, focusing on published spatial proteomics data from a clinically annotated cohort of Cutaneous T Cell Lymphoma patients treated with immune checkpoint blockers, antibodies used to reactivate immune defenses against tumor progression.

Methods. We considered 66 samples from 14 patients, 7 responders and 7 non-responders to anti-PD1 treatment, for which Multiplex ImmunoFluorescence images were produced (Phenocycler by Akoya). 56 antibodies were used to detect the presence of different cell types, including tumor cells, CD4+, CD8+ and Tregs. We used our tysserand and MOSNA libraries for extraction and analysis of tissue networks from the mIF images. Briefly, tysserand enables extraction of networks in which nodes are cells and connections are present if the cells are in contact. MOSNA takes these networks to extract quantitative features that describe the samples' spatial organisation. MOSNA allows exploring local cellular neighbourhoods using the neighbourhood aggregation statistics (NAS) algorithm, which calculates statistical metrics from the attributes of the neighborhood of each cell (e.g. mean, median, variance of marker intensities) creating feature vectors for each cell. Community detection of cells based on these features in a reduced dimensionality space, identifies cellular niches, neighbourhoods with similar composition. This approach results in the definition of niches (proportion of different cell types in each niche) and on the abundance of each niche across samples. We compared our approach to GIOTTO, which quantifies features from spatial datasets using Markov Random Fields. This approach models the conditional dependencies between neighboring cells, assuming that cells close to each other are more likely to belong to the same expression state or cluster.

Results. Applying tysserand and MOSNA we obtained niche definitions and quantifications of niches in each sample, which we used in models to predict patient's response. The first issue we encountered using GIOTTO was that the niches identified in each sample were different, making it impossible to quantify the same niches across samples and produce features for our classification task. We therefore concatenated all sample images together before running GIOTTO to make sure that the same niches could be quantified across the cohort. We then applied a logistic regression to predict patient response based on niche features from GIOTTO and from MOSNA, achieving AUC scores of 0.6 and identifying the niches that had more predictive value.

Perspectives. In the future, we plan to integrate these spatial analysis with immune landscape descriptions based on deconvolution of cell types from bulk RNAseq data in cohorts for which both types of data are available.

PanDelos-plus: A parallel algorithm for computing genetic sequence homology in pangenomic analysis

Simone Colli, Vincenzo Bonnici

University of Parma, Italy

The discovery of homology relations between genes belonging to a set of genomes is a complex challenge in the field of bacterial pangenomics. The aim is to determine how some gene families are distributed among the given genomes. This task requires an all-against-all gene comparison, which makes it computationally tricky and strictly related to the number of analyzed genes. PanDelos is a state-of-the-art approach based on a parameter-free alignment-free homology calculation and gene clustering. In particular, it combines a two-by-two genome comparison by combining a core gene similarity calculation employing k-mer content of genetic sequences with a graph-based unsupervised machine-learning model followed by a paralog-focused refinement procedure for computing the final composition of genetic families. PanDelos shows some of the better performance and reliability of results among the currently available pangenomic tools. However, such an approach is limited in scalability mainly because of the high memory consumption of the core homology detection procedure. Here, we introduce PanDelos-plus, which provides a parallelized solution that makes PanDelos more scalable. We achieve this goal by switching from the original suffix array data structure for k-mer content extraction to a structure more suitable for parallel low-cost homology computation. The proposed structure is gene-centric, meaning each gene is considered the smallest unit to work with. In each two-by-two genome comparison, each gene of one genome is compared to each gene of the other genome. All possible comparison pairs are abstractly represented by a matrix split into independent work units computed separately in parallel. In addition, we introduced some behaviors that allowed us to further optimize the computational process. The first one is based on the representation of k-mers using a unique identifier, thus leaving string processing without compromising their functionality or reliability. The second behaviour is the introduction of a new threshold, which allows us to predict if two genes are different enough to exclude their comparison from the computation. Tests on real and synthetic benchmarks show that PanDelos-plus is much faster and has a lower memory impact than the original version. In particular, we observed an average 4x speed-up of the new approach with respect to the original serial methodology. Such a running time performance increase is proportional to the increase in the number of genes composing the input genomes, independently of the number of genomes. Moreover, we observed a one-tenth average memory consumption reduction compared to the original approach.

Feature Engineering for Protein Sequence Analysis

Gregory Butler

Department of Computer Science and Software Engineering, Concordia University, Montreal, QC, Canada

In this paper, we consider protein sequence analysis using machine learning classifiers. In particular, we consider feature engineering which is a critical step in the development of these classifiers. Protein sequence analysis is a sub-area of bioinformatics and computational biology concerned with computationally determining the structure and function of a protein from its primary structure, the amino acid sequence, and in understanding protein families and their evolution. Our focus has been on transporters which are gates in membranes that organize a variety of vital cellular functions including cell signaling, trafficking, metabolism, and energy production.. Current annotation tools for transporters that predict the substrate for the transport reaction are inadequate. They lag behind tools for other kinds of proteins such as enzymes for metabolic reactions. Most tools for transporters predict the type of substrates, chosen from a small subset of substrate types, without attempting to predict the specific substrate, or predict the family or subfamily for the protein within the Transporter Classification (TC), again without attempting to predict the specific substrate. For systems biology, we must predict each transport reaction; i.e. identify the transport protein and the specific substrate. Our approach to feature engineering leverages diverse types of information as features: (a) composition of the protein sequence in terms of the amino acids and their properties; (b) evolutionary information; (c) positional information for "important" positions (columns) based on conservation, mutual information, or specificity determining positions; (d) regional information in terms of the location within, or near, certain regions of the sequence; (e) sequential information capturing relationships between positions in the sequence. These features may utilize the standard encoding of amino acids, or devise new encodings based on: (i) amino acid properties encoded by reduced alphabets based on size, hydrophobicity, or polarity, or on physico-chemical property values; (ii) exposure of an amino acid (versus hidden) to the environment when the protein is folded. In 2014 the state-of-the-art was TrSSP using 7 substrate classes with an overall MCC of 0.41. Our tool TooT-SC uses evolutionary information via a multiple sequence alignment using TM-Coffee (TMC); positional information from TCS; and pair amino acid composition (PAAC). It applies a Support Vector Machine (SVM) to discriminate each pair of classes. TooT-SC achieves an overall MCC of 0.82. The use of evolutionary information from TM-Coffee on the composition-encoding PAAC improved the MCC by an average of 126.41%. The further use of positional information by filtering out the unreliable columns from the MSA boosted the MCC overall by an average of 128.57%. The cost of classification in TooT-SC is high as each run requires an MSA and the very costly run of TCS.

Context-specific Essential Genes Identification and Prediction by Learning Multi-Omics and Network Data

Maurizio Giordano (1), Lucia Maddalena (1), Mario Manzo (2), Mario Rosario Guarracino (3), Ilaria Granata (1)

(1) Institute for High-Performance Computing and Networking, National Research Council, Naples, Italy
(2) Information Technology Services, University of Naples "L'Orientale", Naples, Italy
(3) Department of Economics and Law, University of Cassino and Southern Lazio, Cassino, Frosinone, Italy

Essential genes are generally defined as necessary for the growth and survival of any organism or cell. Gene essentiality is a key concept in genetics, with implications for basic research, evolutionary biology, systems biology, and precision medicine. In these domains, the task of identifying EGs is challenging due to an increasing demand for procedures able to capture the context-specificity (e.g. tissue, organism or single cells) of gene essentiality. In the context of EG research field, one direction focuses on the development of new computational methods for processing data from gene knock-out experiments. The more harmful the phenotype observed after gene silencing, the more essential the gene. Gene deletion technique becomes complex, costly, labour- and time-intensive when applied at a genome-wide level. Thus, computational approaches are also crucially needed to develop models to predict EGs by learning characteristics of genes that can be associated with their essentiality status. Moreover, these methods allow to compensate for the lack of experimental data due to the inherently limited availability of in vitro models. Prediction models are commonly machine learning models trained in a supervised mode on several genetic characteristics. Some approaches in the literature also exploit physical information automatically learned by deep learning models based on the centrality of genes in protein-protein interaction (PPI) networks. The present work addresses two important tasks concerning context-specific EGs identification, providing novel approaches and related tools: 1) an unsupervised method to identify EGs from gene deletion-derived scores using a binarization scheme based on the Otsu dynamic thresholding algorithm; 2) an ensemble of learners dealing with unbalancing data as the reference machine learning method for EGs prediction based on multi-omics and deep learning features. Multi-omics features involve genomic, transcriptomic, epigenetic, functional, evolutionary and disease-related characteristics gathered from several source database and suitably analyzed and mined. Deep learning features of genes are meant to be network embeddings, i.e. vectorial representation of genes, capturing the centrality of genes in PPI networks according to the centrality-lethality rule: the more central a gene, or its product, the higher its probability of being essential. The methods discussed in in this work are provided as software tools in a unified programming framework, namely HELP (Human Gene Essentiality Labelling & Prediction), and their performance is validated and compared with respect to state-of-the-art methods.

NaStrO: an ultra-rapid, open-source computing pipeline for Nanopore data

Rodolfo Tolloi (1) Marco Prenassi (1) Alberto Cazzaniga (1) Margherita Degasperi (1) Danilo Licastro (1) Stefano Cozzini (1)

(1) Area Science Park

We present Nastro (NAnopore STream Optimized), an ultra-rapid, open-source computing pipeline specifically designed to handle the computational workload of the Oxford Nanopore Technologies' PromethION sequencing device. It operates with exceptional speed for both basecalling and alignment, enabling near real-time data processing and analysis, thereby significantly reducing the latency between sequencing and downstream results to almost none. Moreover, it was developed with an open-source and highly customizable approach, implementing a fair-by-design approach to data management through a easy workflow and an in-depth collection of metadata. ONT has developed an all-in-one software called MinKnow to handle sequencing, basecalling, and alignment for their devices. While this software offers several benefits, it lacks flexibility and is specifically designed for a setup consisting of a sequencing device and a just one single computational unit. Other high-performance pipelines have been also developed, but they often lack the flexibility and ease of updating that Nastro offers: Nastro is built on Jenkins, an open-source automation server, which provides the flexibility to customize and control complex workflows. This adaptability allows for advanced data handling and optimized performance, making Nastro highly versatile. At its core we find ParallelCall, a software developed by us to wrap around ONT's standard basecaller, Dorado, and designed to parallelize the basecalling procedure across multiple computational nodes, using a High Performance Computing (HPC) approach. ParallelCall is higly optimized and can scale the basecalling speed linearly with the computational resources allocated to it. In our tests, using 2 Nvidia DGX nodes from Area Science Park's cluster, Orfeo, we were able to obtain a 16x reduction of the basecalling time compared to the one achieved using the standard ONT computational unit, which is equipped with 4 Nvidia V100 GPUs Once basecalling is completed, each data batch is promptly aligned using MiniMap2. Then, in the final stage of the Nastro pipeline, an introductory analysis is performed, and two detailed reports for each batch are generated—one focused on the basecalling and the other on the alignment. This process provides the user with real-time feedback on their experiment. The pipeline supports an asynchronous data flow, which means that each step operates independently, allowing multiple stream of data to run simultaneously and providing immediate results for each batch of data. The highly customizable and open design of the pipeline enabled the creation of a personalized samplesheet and metadata collection process. This approach allows for seamless integration into a Data Management Plan and implements, by design, a FAIR approach to data handling, ensuring flexible and efficient organization and storage in line with project-specific requirements and objectives.

scVAR: a tool for the integration of genomics and transcriptomics from single cell RNA-sequencing data

Ludovica Celli (1), Samuele Manessi (1), Matteo Barcella (2), Ivan Merelli (1,2)

(1) Institute of Biomedical Technologies, National Research Council of Italy, Segrate (MI), Italy
(2) San Raffaele Telethon Institute for Gene Therapy, Milan, Italy

Single cell RNA sequencing (scRNA-seq) allows the investigation of transcriptomes at single-cell resolution, making this technology a useful tool for studying many diseases in which cell heterogeneity plays a central role. On the other hand, the question of how the genetic variability impacts the transcriptional profile and cell clustering is arising, especially in cancer, as it has a strong genetic component. Currently, the analysis of genetic variations in cancer is done with DNA-sequencing technologies, such as Whole Exome Sequencing (WES) and Whole Genome Sequencing (WGS). However, with the increasing number of scRNA-seq datasets, the urge of developing tools to infer the genomic alterations and the subsequent integration with gene expression is arising. Here, we introduce scVAR, a tool that integrates genomics and transcriptomics data at single cell level. The tool combines gene expression information with single nucleotide variants (SNVs), both detected from scRNA-seq data in three steps: (1) single-cell transcriptome analysis, (2) single cell variants analysis and (3) integrative analysis of the two omics. In step (3) scVAR creates a combined dimensional reduction to accommodate single-cell transcriptomics and genomics data through Deep Learning, using a multiple input autoencoder (AE). The merged omics are processed through the bottleneck of the AE to achieve a combined reduced representation for data in an unsupervised manner. The resulting combined UMAP reduction allows the mapping of each cell multiple times according to the different omics, as well as arithmetical operation between the different points representing the omics. We applied the variants discovery and the scVAR pipeline to scRNAseq samples from diagnoses and relapses of Acute Myeloid Leukemia (AML) and B-cell Acute Lymphoblastic Leukemia (B-ALL). In some cases, the integrated representation was driven by one of the two omics, since clusters resembled those obtained with transcriptomics or genomics only. In other cases, the integrated clusters were able to split the clusters of the starting omics; thus, the integration resolved further heterogeneity. Additionally, peculiar features belonging to diagnoses and relapses of the two diseases came out from the integration. In conclusion, scVAR represent a new technique to analyze transcriptomics data and, coupled with the variants discovery pipeline, returns a new integrated representation describing both omics contributions. Additionally, this method can give useful insights on the intra-cluster heterogeneity due to genetic variability that cannot be investigated with transcriptomics alone.

ELIXIR-IT: the Italian research infrastructure for Life Science data

Graziano Pesole (1,2), and Francesca De Leo (2)

1 Department of Biosciences, Biotechnology and Environment, University of Bari, Via Orabona, 4 70126, Bari, Italy
2 CNR Institute of Biomembranes, Bioenergetics and Molecular Biotechnology (CNR-IBIOM), Via Amendola 122/D-O, 70126, Bari, Italy.

The ESFRI distributed infrastructure for life-science data (ELIXIR), is a unique initiative that brings together Europe's leading life science organisations in managing and exploiting the increasing volume of data being generated by publicly funded research. ELIXIRcoordinates, integrates and sustains bioinformatics resources across European countries and enables users in academia and industry to access and analyze vast amounts of biological data, accelerating scientific discovery.

ELIXIR is organised as a 'Hub and Nodes' model, where the Hub's role is to develop and deliver ELIXIR's scientific strategy coordinating the work done by the Nodes. .

ELIXIR-IT, the Italian node of ELIXIR, plays a pivotal role in fostering bioinformatics research and related translational activities inItaly. The 31 members, coordinated by the CNR, connect Italian researchers with a wealth of computational tools, data resources, and expertise. ELIXIR-IT includes several thematic communities and six operational Platforms: Compute, Data, Interoperability, Omics,Tools, and Training.

ELIXIR-IT aims to enhance industry engagement by promoting partnerships between academia and industry to drive open innovation and address social and data-driven solutions. It aims at offering to the industry partners access to ELIXIR's extensive data, tools resources, computational infrastructure and omics analyses.

Thanks to an investment of approximately 35 million euros by the Ministry of University and Research, ELIXIR-IT has established advanced Omics and ICT facilities. These facilities include state-of-the-art sequencing technologies, high-performance computing resources, and bioinformatics tools. The goal is to provide cutting-edge research capabilities to the scientific community, accelerating discoveries in life sciences.

Among these instruments, top performing sequencing platforms of second and third generation made the ELIXIR Bari Hub one of the most advanced genomic facilities in Europe, also involved in the "Genome of Europe" project for its capable of analyzing a vast number of human genomes in record time.

ELIXIR-IT collaborates closely with industry partners to develop customized bioinformatics solutions that meet their specific needs. By leveraging the expertise of researchers, ELIXIR-IT offers a range of training programs, workshops, and networking opportunities. These initiatives empower researchers and industry professionals to stay up-to-date with the latest advancements in bioinformatics and to connect with like-minded individuals.

By engaging with industry, ELIXIR-IT aims to translate cutting-edge research into practical applications with remarkable societal benefits.

The National Facility for Data Handling and Analysis at Human Technopole: supporting the Italian research community

Roberta Bosotti

Fondazione Human Technopole, Milan, Italy

The National Facility for Data Handling and Analysis (NF-DaHa) at Human Technopole is a new scientific support platform for the Italian research community. It offers advanced data analysis services and supports the development of publicly available tools and resources. NF-DaHa is divided into three units: • Image Analysis (IU1): Provides expertise in image quality control, denoising and restoration, segmentation, and basic quantification. • Omics Analysis (IU2): Conducts standardized analysis of Next-Generation Sequencing data, including single-cell and spatial transcriptomics. • DevOps/Web Development (IU3): Supports the other units by developing analysis pipelines, scientific software tools, and databases. NF-DaHa collaborates closely with other National Facilities at HT, such as the Genomics and Light Imaging facilities, working on internal and external data. It ensures data quality through in-depth quality-control analysis. Primary analysis of omics and image-based datasets will be performed through well-tested, automated pipelines, developed according to industry standards with an emphasis on correctness, robustness, and reproducibility of the results. The pipelines will be extensively documented and shared with the community, to ensure that the analysis processes are sound and meet the requirements of the research projects. Training is a key component of NF-DaHa's mission. The facility not only provides analysis results but also disseminates knowledge to help research groups become independent. Users will be involved at every step in the analysis process through individualized training sessions. Tools and pipelines developed during these processes will be shared. NF-DaHa also organizes public workshops and courses in collaboration with other HT facilities. Another service provided by NF-DaHa is the development and long-term maintenance of scientific software. Often, software developed during research projects is abandoned after the project ends due to lack of funding and resources. NF-DaHa's DevOps unit collaborates with scientists to professionally develop and maintain scientific software and databases, ensuring these resources remain available to the scientific community. Currently in its startup phase, NF-DaHa is offering a limited selection of services in the second half of 2024 and plans to become fully operational in 2025. Recruitment is ongoing for various positions, including bioinformatics scientists, data scientists, image analysis specialists, and software developers. This presentation will describe NF-DaHa's structure, services, and innovative access model designed to maximize its impact on the Italian research community, with a focus on supporting young investigators and those lacking resources for large-scale research. Plans for integrating the facility with national and European bioinformatics networks will also be discussed.

Empowering discovery: advancing life sciences through accessible genomic innovation

Sara Riccardo

NEGEDIA, Italy

Negedia was founded by the Telethon Foundation with the mission of bridging the gap between innovative Next Generation Sequencing (NGS) technologies and their practical applications in life sciences and medical research.

Negedia's purpose is both ambitious and clear: to make cutting-edge technologies accessible across various scientific disciplines, empowering researchers to accelerate discoveries and gain deeper insights into unresolved challenges. Achieving this requires a multimodal approach, where diverse technologies converge to offer a comprehensive, holistic view of the molecular landscape. Negedia highlights a range of advanced technologies, including single-cell sequencing, spatial transcriptomics, mRNA-seq, and genomic applications for DNA studies. This approach demonstrates how integrating leading-edge NGS tools can provide practical, effective solutions to a broad spectrum of biological questions, enhancing researchers' capacity to drive scientific progress.

Empowering Scientific Research: Customized HPC Solutions for Optimal Performance and Efficiency

Paolo Bianco, Marco Fiorletta

I.T.M. INFORMATICA TELEMATICA MERIDIONALE, Via Nuova Poggioreale 11, NAPOLI

This presentation illustrates how our expertise in sizing and customizing high-performance computing (HPC) infrastructures has enabled various research institutions to address and solve complex challenges. We will showcase specific projects developed in collaboration with Dell Technologies, highlighting the value of tailored support. Key benefits include enhanced computational performance, optimized resource utilization, and the flexibility to adapt infrastructure to evolving research demands, fostering innovation and operational efficiency in scientific endeavors.

DietAdhoc® A decision support system for nutrition specialists

Michelangelo Sofo (1) Giuseppe Labianca (2)

(1)Mathematics teacher in high school and IT consultant – Via Malcangi, 141, Trani (Italy)
(2) Nutritionist and health biologist – Via Morrico, 17, Trani (Italy)

The DietAdhoc® system is founded on a mathematical-numerical model based on integer linear programming (P.L.I.) techniques to determine the optimal daily requirement of kilocalories (decisional variables) according to certain constraints on the quantity of macronutrients (carbohydrates, fats, and proteins), micronutrients (fibres, sodium, vitamins, ecc.) and other parameters (glicemic load, ORAC and PRAL). The system has been created tailored for the real working needs of an expert in human nutrition using the human-centered design (ISO 9241-210), therefore it is in step with continuous scientific progress in the field and evolves through the experience of managed clinical cases (machine learning process). DietAdhoc® is a decision support system nutrition specialists for patients of both sexes (from 18 years of age) developed with an Agile methodology which involves an incremental and iterative approach to improving the versions released based on his continuous feedback on the prototypes shown to Dr. Giuseppe Labianca. Its task consists in drawing up the clinical profile of the specific patient by applying two algorithmic optimization approaches (adapted Branch and Bound algorithm and a combinatorial algorithm) on nutritional data and a symbolic solution, obtained by transforming the relational database underlying the system into a deductive database (through queries in Prolog language). The DietAdhoc® system is equipped with a module based on visual data mining techniques for the processing of glycemic curves, histrograms relating to the composition of macronutrients of food portions, and diagrams resulting from the bioimpedance meter SECA mBCA 525. This first version of the system has already been accepted as an application research prototype by the international committee of the BIOSTEC conference, held in Rome from 21 to 23 February 2024.

This is the reference link:
https://drive.google.com/file/d/10IxXmctzVWFLyhmon5XHuwF9Y8Q9mlFP/view?usp=sharing

OASI Biobank: Advancing Asplenia Research through Integrated Bioinformatics and Collaborative Infrastructure.

Marcella Vacca (1), Maurizio Giordano (2), Ilaria Granata (2), Maddalena Casale (3) and Laura Casalino (1)

1. Institute of Genetics and Biophysics “Adriano Buzzati-Traverso”, National Research Council, Naples, Italy

2. Institute for High-Performance Computing and Networking, National Research Council, Naples, Italy

3. Hematology and Oncology Pediatric, Department of Women, Children and General and Specialistic Surgery, University of Campania “Luigi Vanvitelli”, Naples, Italy

The OASI project stems from the collaboration of the Institute of Genetics and Biophysics (IGB) and the Institute of Computing and High-Performance Networks (ICAR) of the National Research Council (CNR) in Naples with the University of Campania "Luigi Vanvitelli," which coordinates the Italian Network of Asplenia (INA). Asplenia refers to the condition resulting from either the surgical removal or an intrinsic dysfunction of the spleen. This condition is associated with various hematologic, oncologic, immunologic, and congenital diseases. Significant complications, such as severe infections and a prothrombotic state caused by coagulation and vascular imbalances, can arise from asplenia, posing life-threatening risks and substantial impacts on long-term disability, public health, and biological costs. Due to the diverse underlying diseases, a comprehensive and systematic approach to asplenia is lacking. Available data are often limited to narrow cohorts of patients with specific diseases, making it difficult to derive reliable information and establish objective clinical guidelines for managing asplenia. The primary objective of the OASI project is to create a Research Biobank for Asplenia, housed at the IGB, that operates on a national scale in the context of the CNR’s Biological Resources Center. This biobank seeks to provide comprehensive and standardized data on asplenia. The OASI Research Biobank will systematically collect, process, and store biological samples from asplenic patients (and healthy volunteers), triangulated with detailed clinical records of the donors available through the INA Clinical Database. By integrating analytical information, including genomic and biochemical data, with clinical data, the OASI Biobank will serve as a research infrastructure and resource hub. These resources will be made accessible under international research standards to facilitate multidisciplinary, multicenter, large-scale observational studies aimed to fill critical gaps in asplenia biomedical research. Furthermore, OASI will promote bioinformatics innovation by leveraging computational biology to provide insights into asplenia, laying the groundwork for future support for the development of therapeutic interventions and clinical decision-making guidelines.

Personalized Cancer Care through Digital Twin Technology: Integrating Patient-Specific Data with Quantitative Systems Pharmacology

Leili Shahriyari

University of Massachusetts Amherst

Our work explores the possibility of creating a digital twin (DT) platform for cancer to better understand the progression of an individual's cancer. By simulating the unique characteristics of each tumor and its response to treatments, we aim to offer insights into personalized cancer care. Our method combines elements of mechanistic modeling, machine learning, and stochastic techniques to develop a DT platform. This platform makes use of diverse data types, such as biological information, biomedical data, and electronic health records (EHR), to create individualized predictions. A central aspect of our approach is the use of a mechanistic model based on quantitative systems pharmacology (QSP). QSP is a computational method used to analyze drug interactions and effects, and it plays a crucial role in our project. We acknowledge that a common challenge in QSP modeling is accurately determining parameters, especially since traditional models often assume a general uniformity across different patients' diseases. This assumption can lead to limitations when calibrating parameters using varied data sources. Our objective is to build a more personalized DT by concentrating on individual patient data for parameter estimation. We adjust the QSP model parameters for each patient based on their unique data. Through detailed sensitivity analysis and uncertainty quantification, we identify key interactions in the model and define the range of our predictions. By integrating this tailored QSP model with insights into cellular and molecular interactions, we hope to better predict how cancer evolves and responds to specific treatments. We are excited about the potential this has for advancing personalized cancer therapy, though we are aware of the challenges and complexities involved in this endeavor.

inDAGO: a user-friendly graphical interface for dual RNA-seq data analysis

Carmine Fruggiero (1,2), Gaetano Aufiero (2), Nunzio D'Agostino (2)

(1) Department of Electrical Engineering and Information Technology, University of Naples Federico II, Via Claudio 21, 80125 Napoli, Italy

(2) Department of Agricultural Sciences, University of Naples Federico II, via Università 100, 80055, Portici, Italy

Transcriptomics studies of host-parasite interactions present significant challenges, particularly when the contact region between the two interacting organisms is difficult to isolate. Dual RNA-seq analysis offers a powerful alternative by allowing researchers to separate mixed transcripts in silico, providing a comprehensive view of transcriptome reprogramming in both host and parasite during infection. However, many existing pipelines for dual RNA-seq require a strong understanding of programming languages, as they are often distributed with command line interfaces in R or Linux environments. This creates a barrier for researchers without a programming background, highlighting the growing need for user-friendly software that can make these analyses more accessible to a broader scientific audience. Here, we introduce inDAGO, an open-source, cross-platform graphical user interface designed for dual RNA-seq analysis, offering two distinct approaches: sequential and combined. Built on the R Shiny framework, inDAGO is organized into modular components, the best available software for data processing alongside custom-developed R scripts for specific steps. Additionally, inDAGO includes a module for traditional bulk RNA-seq analysis. The interface is designed to be user-friendly, catering to researchers with varying levels of technical expertise, without compromising on the precision and flexibility needed for complex analyses. Users can easily customize a wide range of parameters to tailor the analysis to their specific dataset. Ultimately, inDAGO guides users through the entire process - from initial data input to to the identification of differentially expressed genes (DEGs) -while providing intermediate result and publication-ready outputs at the conclusion of each module.

Modeling Variations in Antibody Response Magnitude and Longevity

Paola Stolfi (1), Filippo Castiglione (2), Enrico Mastrostefano (1), Giovanni Messuti (3), Luca Pugliese (4), Silvia Scarpetta (3), Antonella Prisco (4)

(1) Institute for Applied Computing, CNR, Italy

(2) Technology Innovation Institute, United Arab Emirates

(3) Department of Physics, University of Salerno, Italy

(4) Institute of Genetics and Biophysics, CNR, Italy

Long-lasting antibody responses are pivotal for both protective immunity and autoimmunity. Yet, the intricate mechanisms that dictate the duration of these responses remain only partially elucidated. By employing an agent-based in silico model, we simulated the generation of short-lived and long-lived plasma cells during the immune response to an adenoviral COVID-19 vaccine, postulating that antigen-specific plasma cells have a certain probability of attaining an extended half-life. This hypothesis implies that the quantity of antigen-specific plasma cells generated in the initial immune response, coupled with their likelihood of becoming long-lasting, influence the magnitude of the antibody response months after immunization. Interestingly, our simulations unveiled two distinct clusters among individuals several months post-vaccination, delineating markedly divergent dynamics in antibody titers: one group exhibited sustained elevated antibody levels (sustainers), while another witnessed a decline (decayers). Notably, the absence of long-lived plasma cells in the decayers distinguished them from the sustainers. Leveraging machine learning clustering on antibody titers, we achieved an accuracy of 0.925 in identifying the decayers 28 weeks following the initial dose.In this in-silico system, the difference between sustainers and decayers stems from stochastic inter-individual differences in the immune repertoire and the efficacy of priming. We speculate that, in real life, aged and immunocompromised people may be prone to the decayer pattern and may benefit from receiving their vaccine booster after a shorter interval. We are comparing our model's predictions with clinical data on the antibody response to SARS-CoV-2 Nucleoprotein and Spike post-COVID-19 infection or vaccination to validate and refine our model. Specifically, we are harnessing machine learning methodologies on data sourced from published immunological studies to discern patterns in the dynamics of the antibody response.

FUNDING

The authors acknowledge support by the following grants: European Union – Next Generation EU. PRIN P2022BS4EH Machine-learning analysis of the dynamics of the antibody response induced by Covid-19 vaccination (to AP and SS). CNR project FOE-2021 DBA.AD005.225 (to AP). Next Generation EU-MUR PNRR Extended Partnership initiative on Emerging Infectious Diseases PE13 INF-ACT (to AP). CNR project InvAt- Invecchiamento attivo e in salute (FOE 2022) (to AP).

Metagenomic analyses identify biosynthetic gene clusters of Mediterranean sponges leading to bioactive products

Roberta Esposito (1), Serena Federico (1,2), Michele Sonnessa (3), Sofia Reddel (3), Marco Bertolino (2), Marco Giovine (2), Marina Pozzolini (2), Valerio Zupo (4), Maria Costantini (1)

(1)Department of Ecosustainable Marine Biotechnology, Stazione Zoologica Anton Dohrn, Via Ammiraglio Ferdinando Acton, n. 55, Napoli, Italy
(2) Department of Earth, Environmental and Life Sciences, University of Genoa, Genoa, Italy
(3) Bio-Fab Research Srl, Rome, Italy
(4) Department of Ecosustainable Marine Biotechnology, Stazione Zoologica Anton Dohrn, Ischia Marine Centre, Naples, Italy

The oceans cover over 70% of our planet, hosting most of the diversity of our biosphere. The critical chemical and physical conditions of the sea, impacted by several human activities, favoured the production of a remarkable variety of novel molecules by marine organisms. These compounds are quite interesting in terms of diversity as well as structural and functional features that could be important for biotechnologies, as compared to molecules isolated from terrestrial organisms. Marine sponges and in particular those in the class Demospongiae, are among the most promising organisms producers of bioactive compounds, also because they provide shelter for a wide variety of associated microorganisms. In fact, they consistently harbour dense and diverse microbial communities. Interestingly, many microorganisms are specific to the host sponges. Such microbes, which include bacteria, archaea, and unicellular eukaryotes (fungi and microalgae), contribute up to 40% of the sponge volume and have remarkable effects on the host biology, being a source of biologically active compounds, having possible pharmaceutical and cosmeceutical applications. An ongoing debate is devoted to define if secondary bioactive metabolites are synthetized by the cells of sponges, by the associated microorganisms or by the interaction between microorganisms and sponges. On this line, the characterization of the bacterial communities associated with sponges becomes crucial. In addition, the laboratory (and/or in situ) cultivation of sponges is still considered a challenge. Therefore, the application of environmentally-friendly approaches for the identification of new biosynthetic gene clusters, using metagenomics analyses, is a good alternative to the over-utilization of marine resources and to destructive collection practices. Research on such Mediterranean sponges as Agelas oroides, Haliclona vansoesti and Geodia cydonium showed that they were good candidates for isolation of new bioactive compounds, to be proposed as alternative drugs for treatment of solid tumour. Furthermore, metataxonomic analyses revealed a great variability of the host-specific microbial communities in these three sponges. Here, we investigated the microbiomes composition and its metabolic potential in A. oroides, H. vansoesti and Geodia cydonium. Metagenomic analyses were performed, describing the genes involved in metabolic synthesis of such bioactive compounds as fatty acids, antibiotics, vitamins, terpenoids, sterols and pigments. The application of new metagenomic techniques provides tools to explore the metabolic diversity of marine microbes. It facilitates the characterization of the genomic repertoire of sponge-associated microorganisms at an unprecedented resolution. In addition, it provides insights into the molecular mechanisms underlying microbial-sponge interaction.

Quality of semi-automated de novo genome assembly

Aleksandra Swiercz (1,2), Artur Laskowski (1), Alicja Dzik (1), Pawel Wojciechowski (1,2), Piotr Lukasiak (1), Jacek Blazewicz (1)

(1) Institute of Computing Science, Poznan University of Technology, Poland
(2) Institute of Bioorganic Chemistry, Polish Academy of Sciences, Poland

The first draft of the human genome project was published in year 2000. It was incomplete, especially in highly repetitive regions like centromeres or telomeres. With the development of third-generation technology, it has become possible to fill most of the gaps in the reference genome (T2T consortium). However, long-read technologies can cover large portions of the genome, this involves higher error rates. Thus, de novo sequencing projects often use different sequencing technologies to, on the one hand, obtain long DNA fragments and, on the other hand, improve the sequence quality with short reads. In our project Genomic Map of Poland we sequenced a cohort of 5000 Polish citizens to check the population diversity. We constructed also a diploid reference genome using de novo assembly pipeline based on the trio: mother, father and child. To complete this task, several sequencing technologies were used in the pipeline: short reads, PacBio HIFI, Hi-C and ultra-long Nanopore. The resulting chromosome-wide scaffolds were compared to reference genomes (GRCH38 and CHM13). We then assessed the quality of our reference genome assembly by analyzing the consistency of k-mers appearing in the genome sequence and in short and long reads. It was also checked whether k-mers specific to only one of the parents (mother or father) occur in the copy of the chromosome inherited from one parent (each copy of the chromosome contains small differences in the sequence, specific to the individual). Various k-mer features indicate the high quality of the assembled genome.

Deep discriminative models in the detection of amyloid signaling motifs

Jakub Gałązka, Krzysztof Pysz, Witold Dyrka

Politechnika Wrocławska, Wydział Podstawowych Problemów Techniki, Katedra Inżynierii Biomedycznej, Poland

Amyloid motifs are very short domains (typically around 25 amino acids in length) that facilitate protein aggregation into a polymeric fibrillary structure or the amyloid fold. Best known for their role in pathological amyloidoses, such as Alzheimer's disease, amyloid motifs are also involved in many physiological functions, including biofilm formation, oligomerization and signal transduction. Among others, amyloid signaling participates in the innate immune response in fungi, bacteria and archaea. The mechanism involves signal transduction from the receptor protein to the effector protein by imposing an amyloid fold and oligomerization. A dozen families of bacterial and fungal amyloid signaling sequences that occur at the N-termini of NOD-Like Receptors and at the C-termini of cooperating effectors have been identified and described, yet the goal is to characterize the entire signaling amyloidosome.

Existing computational tools for identification of short amyloidogenic hotspots lack specificity for meaningful searches for signaling amyloid motifs in entire genomes. On the other hand, profile-based models are effective only when trained for particular motif families. To address these difficulties, we developed and thoroughly compared three generalized models of amyloid signaling motifs based on computational linguistic modeling and deep learning approaches. The models were trained on a set of variable-length bacterial C-terminal amyloid signaling sequences from ten diverse families as the positive sample and protein fragments representing the general protein space as the negative one.

Overall, the best performance was achieved by the finetuned ProteinBERT model based on the Transformer neural network architecture, which maintained the recall of almost 70% at FPR of 10e-5 for N-terminal bacterial motifs and the recall of ~50% at FPR of 10e-4 for fungal amyloid motifs. We also extracted the embeddings of predicted sequences to visually represent their relative placement in two-dimensional space using the UMAP algorithm. Visualizations extracted from the Transformer model suggest that ProteinBERT recognizes families as both separate from the negative dataset and distinct from other families. Our work provides further evidence for the usefulness of deep learning models in protein analysis and classification, and highlights the potential of these methods in solving needle-in-a-haystack bioinformatics tasks, such as amyloid motif searches in entire genomes. Moreover, a better understanding of microbial amyloid signaling could help combat amyloidoses, as microbiome amyloids interact with human proteins.

AI-Driven Antibiotic Resistance Prediction in Hospital and Clinic Settings

Marco Benedetto (1,5), Antonio Facchiano (2), Giuseppe Piccinni (2), Cristina De Leo (2), Luciano Sobrino (2), Annarita Panebianco (2), Francesco Facchiano (3), Angelo Facchiano (4), Stefano Tagliaferri (5), Michele La Rocca (6), Roberto Tagliaferri (1)

(1) Department of Management and Innovation Systems, University of Salerno, Fisciano (SA) 84084, Italy
(2) Istituto Dermopatico dell'Immacolata, IDI-IRCCS, Rome, Italy
(3) Department of Oncology and Molecular Medicine, Istituto Superiore di Sanità, Rome, Italy
(4) Laboratory of Bioinformatics and Computational Biology, Institute of Food Science, CNR, via Roma 52 A/C, 83100 Avellino, Italy
(5) Kelyon S.r.l., via Benedetto Brin, 59 C5/C6, 80100 Naples, Italy
(6) Department of Economics and Statistics, University of Salerno, Fisciano (SA) 84084, Italy

The accurate selection of antibiotic treatment is critical in the face of rising bacterial resistance, which poses a significant global health threat. AI-driven clinical decision support systems offer a ground-breaking advantage over empirical methods reliant on culture-based testing, which is both time-consuming and labor-intensive. In this context, machine learning presents a promising approach to address the issue of antibiotic resistance. This study aims to develop and train machine learning and deep learning models to predict antibiotic susceptibility, offering faster and more accurate treatment recommendations. Data for the study were collected over five years (2018-2022) from two sites in the Lazio region of Italy. The dataset consists of 10551 isolates collected from 6992 inpatient and outpatient patients. The population had a mean age of 60.4 ± 21.3 years, with a predominance of female subjects (56.7%). Several machine and deep learning models were trained to predict antibiotic susceptibility test results, including logistic regression, CART (Classification And Regression Tree), random forest, XGBoost (eXtreme Gradient Boost), and TabNet. The datasets include attributes such as patient demographics, sample details, department details, Gram stain classification, bacterial species, previous bacterial infections, 33 antimicrobial substances, and their corresponding antibiotic susceptibility results and previous resistances. Since Gram staining results are available before full bacterial identification, models were trained using partial and complete information about the detected organism. By integrating susceptibility data with biochemical test results, a subset dataset containing 990 observations was created. Multiple feature sets from this dataset were explored for model training. The classifiers were compared using a 10-fold cross-validation approach, and their performances were evaluated using an 80:20 split. Models were evaluated using the area under the receiver operating characteristic curve (AUROC), with XGBoost demonstrating the highest discriminative performance. When only Gram stain was available, the XGBoost model trained on the Gram-positive dataset achieved an AUROC of 0.862, while the model trained on the Gram-negative dataset achieved 0.843. For models trained on a subset of isolates positive to a known bacterial species, XGBoost achieved AUROC scores of 0.887 for E. coli, 0.912 for E. faecalis, 0.903 for K. pneumoniae, 0.884 for P. aeruginosa, 0.875 for P. mirabilis, and 0.889 for S. aureus. Incorporating biochemical features further enhanced performance, with the XGBoost model reaching an AUROC of 0.914. These findings indicate that machine learning models like XGBoost can effectively support clinical decision-making by predicting antibiotic susceptibility. If integrated into clinical decision support systems, AI-driven models can improve treatment selection and mitigate the impact of antimicrobial resistance (AMR).

Classification and biochemical evaluation via Raman and Surface-enhanced Raman scattering spectroscopy of breast cancer cell lines expressing different levels of HER2

Alessandro Esposito (1), Sara Spaziani (2,3), Giovannina Barisciano (4), Giuseppe Quero (5), Manuela Leo(5), Vittorio Colantuoni (5), Maria Mangini (1), Marco Pisco (2,3), Lina Sabatino (5), Anna Chiara De Luca (1), Andrea Cusano (2,3)

(1) Institute for Experimental Endocrinology and Oncology G. Salvatore, IEOS, second unit, 80131, Naples, Italy
(2) Centro Regionale Information Communication Technology (CeRICT Scrl), 82100, Benevento, Italy
(3) Optoelectronic Division-Engineering Department University of Sannio, 82100, Benevento, Italy
(4) Department of Sciences and Technologies, University of Sannio, 82100, Benevento, Italy
(5) Biosciences and Territory Department, University of Molise, 86090 Pesche, Italy

HER2 (Human epidermal growth factor 2) enriched breast cancer (BC) screening is pivotal in clinics. Still, the diagnosis remains too uncertain, with at least 20% of incorrect diagnoses1. Additionally, diagnostic standard techniques, such as Immunohistochemistry (IHC) and Fluorescence Hybridization in situ (FHS) require active staining and elevated costs of reagents2. In contrast, Label-free techniques are reagent-free and inexpensive. Specifically, Raman spectroscopy (RS) exploiting the Inelastic-scattering properties of the light, is a label-free multiplexing technique, that can retrieve the chemical composition of biological samples3. Thus, BC cells are classified by the relative differences in the quantity of specific macromolecules, since a correlation between the levels of aromatic amino acids and the level of lipids and phospholipids has been attributed to the different aggressiveness and invasiveness rates of tumor cells 4. But nevertheless, RS suffers from a lack of sensitivity and doesn't achieve mass spectroscopy or liquid chromatography performances enabling biomarkers quantification. Hence, RS is combined with specific plasmonic nanostructures that dramatically improve its sensitivity and it is usually referred to as Surface Enhanced Raman Scattering (SERS) Spectroscopy, reaching even a single molecule sensitivity5. Using both RS and SERS, we verified the HER2 cellular concentration and the biochemical composition of BC cells. Operatively, we have selected different BC cell lines, expressing different levels of HER2 directly related to the BC prognostic outcome. Specifically, we have chosen stable clones from BC cell lines: MCF10A (healthy cells), SKBR3 (HER2-positive), MDAMB443 (triple Negative, malignant), Sh-SKBR3 (HER2 silenced clones from SKBR3), that have been characterized via fluorescence-activated cell sorting (FACS) and Western blot (WB) to assess the homogeneity and the proper expression of HER2 protein. Finally, BC cell Raman spectra have been classified with principal component analysis (PCA) and linear discriminant analysis (LDA) analysis to evaluate their biochemical differences in lipids and aromatic amino acids content and correlate them with different HER2 levels, obtained by SERS analysis. Thus, we obtained high-accuracy discrimination of the BC cell Raman signal due to these specific constituents and precise HER2 quantification by SERS, paving the way to single-cell Raman screening of HER2 as a cheap and complementary tool to the standard diagnostic techniques, once validated on tissue biopsies.

Enhancing predictions of protein stability changes induced by single mutations using MSA-based language models

Francesca Cuturello (1), Marco Celoria (2), Alessio Ansuini (1) e Alberto Cazzaniga (1)

(1) Area Science Park, Trieste, Italy
(2) CINECA, Bologna, Italy

Protein Language Models offer a new perspective for addressing challenges in structural biology, while relying solely on sequence information. Recent studies have investigated their effectiveness in forecasting shifts in thermodynamic stability caused by single amino acid mutations, a task known for its complexity due to the sparse availability of data, constrained by experimental limitations. To tackle this problem, we introduce two key novelties: leveraging a protein language model that incorporates Multiple Sequence Alignments to capture evolutionary information, and using a recently released mega-scale dataset with rigorous data preprocessing to mitigate overfitting. We ensure comprehensive comparisons by fine-tuning various pretrained models, taking advantage of analyses such as ablation studies and baselines evaluation. Our methodology introduces a stringent policy to reduce the widespread issue of data leakage, rigorously removing sequences from the training set when they exhibit significant similarity with the test set. The MSA Transformer emerges as the most accurate among the models under investigation, given its capability to leverage co-evolution signals encoded in aligned homologous sequences. Moreover, the optimized MSA Transformer outperforms existing methods and exhibits enhanced generalization power, leading to a notable improvement in predicting changes in protein stability resulting from point mutations. Code and data are available at https://github.com/RitAreaSciencePark/PLM4Muts.

Multi-omics data integration methods for cancer-subtyping, drug discovery and tumor-model alignment

Aurora Brandi (1), Massimiliano Romano (1), Luigi Ferraro (2), Barbara Majello (1), Michele Ceccarelli (2), Giovanni Scala (1)

(1) University of Naples Federico II, Italy

(2) University of Miami, FL, USA

Background: Cancer is a highly heterogeneous collection of complex diseases, traditionally classified based on their tissue of origin and histological features. However, the significant plasticity and molecular heterogeneity of cancer cells can lead to marked variations in clinical outcomes and treatment responses within the same tumor type, making necessary the identification of cancer subtypes that reflect as much as possible the complex tumors molecular setup. This diversity also complicates the selection of laboratory models (e.g., tumor-derived cell lines) that effectively recapitulate the biology of specific cancer types (subtypes). Identifying models that closely reflect the altered molecular background of a patient's tumor is crucial to enhance the translational potential of in vitro findings, particularly in drug screening studies. Methods: Here, we introduce three multi-omics computational tools for: (i) cancer-subtyping, (ii) tumor-model alignment, and (iii) cancer-specific drug repositioning. The first tool, MultiOmics Network Embedding for SubType Analysis (MoNETA), enables the fast and scalable identification of relevant multi-omics relationships between biological samples. The second tool, MultiCelligner, performs the alignment of tumor and cell line models based on multiple omics layers, enabling the identification of the most relevant cell line model recapitulating a given tumor (sub-)type in terms of single or multiple molecular districts conformation. The third tool, multi-omics genes and Drugs Network Embedding (MiDNE), predicts multi-omics associations between genes and drugs by constructing multiomics gene-centric networks which are then integrated with targeting drugs. Conclusions: MoNETA, MiDNE and MultiCelligner are three stand-alone tools that, when used together, offer a comprehensive framework for identifying patient specific cancer subtype, selecting the most biologically relevant cell lines for specific cancer subtypes and predicting potential therapeutic compounds.

Integrative analysis of heterogeneous high-throughput transcriptomic data for promoter selection in bacterial genomes to support microbial synthetic biology

Debora Dallera (1), Stefano Quaranta (2), Davide De Marchi (1), Paolo Magni (1), Lorenzo Pasotti (1)

(1) Department of Electrical, Computer and Biomedical Engineering, and the Centre for Health Technologies, University of Pavia, Pavia 27100, IT
(2) the Department of Biology and Biotechnology, University of Pavia, Pavia 27100, IT

Promoters are fundamental elements in synthetic biology, serving as the key regulatory components that enable the design of sophisticated sense-and-respond functions, such as biosensors, in engineered bacterial strains. While our understanding of non-model bacteria continues to expand, the availability of promoter libraries exhibiting a wide range of transcriptional activities and diverse induction patterns remains notably scarce. This limitation poses significant challenges for researchers seeking to engineer bacterial systems with specific, desired functionalities. One promising data-driven avenue to address this challenge lies in the analysis of high-throughput transcriptomic data, which are increasingly generated and made publicly available. These datasets hold valuable insights into gene expression profiles across various conditions; however, the inherent heterogeneity of public data presents obstacles to effective joint analysis. The data may derive from multiple experimental technologies (e.g., microarray and RNA-Seq), diverse strains, and numerous bacterial species, complicating the process of comparing and integrating findings. To tackle these challenges, we present a novel bioinformatics pipeline, implemented in R, specifically designed to harmonize and jointly analyze transcriptomic data from multiple sources. It facilitates the selection of genes characterized by target expression means and coefficients of variation, thereby enabling the identification of both constitutive and inducible promoters. These promoters are essential for the development of biosensors that can respond appropriately to environmental stimuli, making our pipeline a valuable asset in synthetic biology research. Several key modules are included, each serving a distinct function. The pipeline includes capabilities for automated retrieval of public transcriptomic data for selected species, processing of expression data, operon identification, and the crucial harmonization of gene IDs. This ID harmonization is vital for enabling multi-strain analyses and facilitating comparisons across related species that may share similar regulatory patterns. By overcoming the challenges posed by batch effects and variability in data generation, our pipeline is expected to enhance the reliability of promoter identification. We applied this tool to case studies focusing on promoter selection in different bacterial species, demonstrating its practical utility and effectiveness in identifying promoters with robust activities across conditions, or promoters with highly variable activity that may serve as candidates for biosensor design. This tool is complementary to existing methods and is poised to significantly contribute to the informed design of sense-and-respond functions in engineered bacterial systems. By facilitating the identification of appropriate promoters with an automated data-driven approach, we expect our tool to support advances in the field of synthetic biology.

Machine learning and explainable AI for transcriptomic analysis in Multiple Sclerosis

Silvia Giulia Galfrè, Francesco Massafra, Samuele Punzo, Corrado Priami, Alina Sirbu

University of Pisa, Italy

Multiple sclerosis (MS) is a chronic autoimmune disease characterized by the immune system's attack on the central nervous system, causing neurological damage. Computational biology and single-cell technologies offer new approaches to explore the genetic and cellular mechanisms underlying this condition. This work presents a computational analysis to identifying genetic markers in multiple sclerosis (MS) using advanced machine learning techniques with microarray and single-cell RNA sequencing (scRNA-seq) data on peripheral blood mononuclear cells (PBMC) samples and cerebrum spinal fluid (CSF) - only scRNA-seq. The study employed specific steps of data preprocessing, including dataset integration and normalization. We applied RMA, Combat and MinMax for the microarray datasets and CCA dataset integration (using Seurat) for the scRNA-seq datasets. Subsequently, machine learning models were trained to classify samples into MS and control. For scRNA-seq datasets multiple models were obtained, for different cell populations and tissue types. The analysis focused on Treg, Th1 and Th17 cell subpopulations. For microarrays, instead, we explored models obtained with different data integration techniques. The models obtained were studied with explainable AI (XAI) tools, such as SHAP (SHapley Additive exPlanations) and ablation studies, to highlight the most important genes. This methodology can complement standard differential analysis tools and underline different biomarkers and their interactions. Key genes, such as DDX3Y, HLA-DRB1, and USP25, were found to be important in sample classification for microarray datasets. In the scRNA-seq analysis, comparative analyses between tissues highlighted specific gene markers enriched in cell clusters. With the SHAP approach, three genes were detected as commonly important: DCAF12 a gene involved in T-cell activation and protein ubiquitination, TRAF4 also involved in ubiquitination and with the immune system, SCL40A1 a protein that can be ubiquitinated and is involved in lymphocyte homeostasis. These studies can give indications regarding the molecular mechanisms underlying MS. The identified genes, many of which are associated with immune responses and inflammation, suggest potential targets for therapeutic intervention, to be further evaluated and validated in laboratory experiments.

Mathematical Modeling of Phage-Mediated CRISPRi System for Inhibiting Antibiotic Resistance

Massimo Bellato (1,2), Chiara Cimolato (2,3), Sara Letrari (1,2), Luca Schenato (2,3)

(1) Dept. of Molecular Medicine, University of Padova, Padova, Italy
(2) "Centro Studi di Biologia Sintetica - SynBio UniPD", University of Padova, Padova Italy
(3) Dept. of Information Engineering, University of Padova, Padova, Italy

Antimicrobial resistance (AMR) poses a significant challenge to global health. One promising strategy to combat this issue involves the use of engineered non-lytic phages to transfect antibiotic-resistant bacteria with a CRISPR interference (CRISPRi) system, with the goal of restoring their susceptibility to antibiotics. To explore this approach and support the rational design of these engineered bio-therapeutics, we have developed a mathematical model that captures the fundamental mechanisms associated with phage infection, CRISPRi delivery, and mutation dynamics. The model comprises three interconnected components: bacterial growth, phage transfection dynamics, and mutation occurrences. Bacterial growth is represented using a chemostat-like abstraction, where nutrient availability and washout rates influence population size. The phage-mediated delivery of the CRISPRi system is modeled using delayed differential equations (DDE) to account for the time delay in CRISPRi production following infection. Specifically, transit compartmental models (TCM) are employed to depict various stages of bacterial infection and CRISPRi system activation, acknowledging the possibility of mutations during this interval. Mutation dynamics are incorporated into the model to reflect alterations in both bacterial DNA and phagemid-carried CRISPRi systems. The model considers multiple bacterial states, including those with functional or mutated CRISPRi circuits, while tracking the population of antibiotic-resistant bacteria that might evade treatment. Simulations were conducted to assess a range of initial conditions, including the number of phages and the rates of antibiotic-induced killing, identifying critical thresholds for effective treatment where an adequate supply of phages and antibiotic action significantly diminishes bacterial survival and evasion rates. Mutational analysis indicates that the emergence of mutants, particularly within the CRISPRi system or bacterial targets, is significant. Nevertheless, the model also shows that with proper dosing of phages and antibiotics, the population of surviving resistant bacteria can be significantly reduced. These findings suggest that the phage-CRISPRi system could be a helpful way to tackle antibiotic resistance, as long as the right treatment guidelines are set up to reduce the chances of developing resistance due to mutations. This model serves as a valuable framework for optimizing phage-CRISPRi therapies and enhancing our understanding of the dynamics involved in reversing bacterial resistance. Future studies will aim to refine the model by incorporating additional complexities, such as spatial constraints and varying mutation rates, and also comparing the conceived solution with other therapeutic strategies such as CRISPRi coupled with bacterial conjugation or lytic-phage-based therapy.

ABSTRACTS OF POSTERS

Anti-cancer compound biosynthetic pathways in the marine diatom *Cylindrotheca Closterium*

Nova Yurika (1,2), Eleonora Montuori (1,3), Chiara Lauritano (2)

(1) Marine Biology Research Group, Ghent University, Krijgslaan 281, B-9000 Gent, Belgium
(2) Ecosustainable Marine Biotechnology Department, Stazione Zoologica Anton Dohrn, Italy
(3) Department of Chemical, Biological, Pharmaceutical and Environmental Science, University of Messina, Italy

Marine microalgae are attracting huge interest for the isolation of novel natural compounds with possible industrial applications, including the pharmaceutical sector (Saide et al., 2021). However, these compounds are often present in small amounts. Many research efforts are focusing on how to increase the production of the compounds of interest. It is known that nutrient modification could trigger changes in microalgae metabolism and influence the production of possible bioactive molecules. In the current study, we used two different nitrogen sources, nitrate and urea, to cultivate the diatom Cylindroteca closterium. Our aim was to identify the transcripts encoding enzymes involved in the synthesis of the well-known anticancer compounds 1,5 Aminolevullinic acid (5-ALA) and Pheophorbide a (PPBa), and analyze their expression levels in the two nutrient conditions. We cultivated and harvested C. closterium in exponential and stationary phase for both nutrient conditions. By mining the already published transcriptome of C. closterium (Elagoz et al., 2020) and by using internal blast, we looked for the transcripts encoding enzymes involved in the synthesis of the anti-cancer compounds. In particular, we designed primers and performed reverse-transcription real time PCR for eleven genes. Results showed that urea medium treatment induced a general trend of increased expression levels of all the gene of interest, with cells cultured in the stationary phase showing higher gene expression compared to the exponential phase. Chemical analyses will confirm if this is the condition with highest quantity of the anticancer compounds. This study provides new insights into the culturing conditions favoring the release of bioactive compounds, and in the identification of candidate genes for genetic engineering experiments to implement the production of anticancer drugs.

References
-Elagoz, A.M., Ambrosino, L., Lauritano, C., 2020. De novo transcriptome of the diatom Cylindrotheca closterium identifies genes involved in the metabolism of anti-inflammatory compounds. Sci Rep 10, 4138. https://doi.org/10.1038/s41598-020-61007-0
-Saide, A., Martínez, K.A., Ianora, A., Lauritano, C., 2021. Unlocking the Health Potential of Microalgae as Sustainable Sources of Bioactive Compounds. IJMS 22, 4383. https://doi.org/10.3390/ijms22094383

Interpretable machine learning models for pathogenicity prediction

Javier Guerrero-Flores (1,2)*, Natàlia Padilla (1)*, Xavier de la Cruz (1,3)

1 - Vall d'Hebron Institute of Research (VHIR), Universitat Autònoma de Barcelona, Barcelona, Spain
2 - Universidad de Barcelona, Barcelona, Spain
3 - Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain
**These authors contributed equally to this work.*

Artificial Intelligence (AI) plays a pivotal role in the digital transformation of various fields, significantly impacting Genomic Medicine by driving foundational changes. However, despite the promise, advancements in this area have experienced a slowdown in recent years, primarily due to unresolved challenges associated with the experimental techniques underpinning Genomics. A notable issue is the Variant Interpretation Problem, which is central to clinical Next-Generation Sequencing (NGS). This problem poses a straightforward yet critical question: can we determine whether a genetic variant identified through NGS is pathogenic or benign? Currently, several AI-based tools, known as pathogenicity predictors, aim to provide answers to this question. However, their applicability in the clinical setting is limited by their black box nature, which renders them difficult to interpret. The research presented focuses on overcoming the interpretability challenge to effectively apply AI to clinical and biomedical challenges using genetic variants information. Specifically, we are pioneering a new, interpretable pathogenicity predictor leveraging the advanced algorithm FasterRisk, a machine learning tool devised by Cynthia Rudin's team. This tool aims to produce prediction scores easily understandable by users. Our poster details the development, training, and evaluation of this predictor, with a focus on comparing its performance against conventional, black box AI tools. In particular, we will show how our results validate, in our field, Rudin's hypothesis that interpretable methods can match or surpass the effectiveness of traditional black box approaches. Given the innovative nature of our work, we will also discuss interpreting the outcomes of our predictor through a series of illustrative examples, thereby illuminating its practical use.

Assessing the impact of sailing on motor functions and quality of life in young patients with rare skeletal disorders: a quantitative approach

Davide Scognamiglio (1), Manila Boarini (1), Giuseppina Mariagrazia Farella (2), Giacomo Villa (3,4), Enrica Di Sipio (3), Giulia Rogati (5), Lisa Berti (2,6), Silvana Sartini (7), Alberto Leardini (5), Luca Sangiorgi (1)

(1) Department of Rare Skeletal Disorders, IRCCS Istituto Ortopedico Rizzoli, Bologna, Italy

(2) Physical Medicine and Rehabilitation Unit 1, IRCCS Istituto Ortopedico Rizzoli, Bologna, Italy

(3) Euleria srl Società Benefit, Rovereto, Italy

(4) Department of Electronics, Information and Bioengineering (DEIB), Politecnico di Milano, Milano, Italy

(5) Movement Analysis Laboratory, IRCCS Istituto Ortopedico Rizzoli, Bologna, Italy

(6) Department of Biomedical and Neuromotor Sciences, Alma Mater Studiorum, University of Bologna, Bologna, Italy

(7) Physical and Rehabilitation Medicine Unit 2, IRCCS Istituto Ortopedico Rizzoli, Bologna, Italy

Rare skeletal disorders can significantly impair balance, posture, and joint mobility, often causing lifelong disabilities and an overall lower quality of life. Quantifying these features remains challenging. The aim of this study was to evaluate proprioception, posture, and arms mobility, as well as quality of life and perceived health status in young patients with rare skeletal disorders. Eight young patients affected by Multiple Osteochondromas, Ollier Disease, and Osteogenesis Imperfecta have been enrolled in this study and data were collected from two different sources: inertial movement units (IMUs) and patient-reported outcomes (PROs). We assessed these metrics before and after a three-day sailing experience. As for IMUs data, we computed single metrics like sway and root mean square. In order to integrate multiple metrics and sensors together, a Composite Proprioception Score (CPS) was calculated by measuring the standard deviations of both the Center of Pressure (COP) and trunk positions during a static balance test. Posture was evaluated by computing a Postural Score (PS) which is based on the mean average euclidean distance between the trunk and COP. These scores provided an objective measurement framework for analyzing the impact of the sailing experience on these young patients' balance. Exploratory data analysis was performed on PROs data to identify changes in health-related quality of life, psychological well-being, self-esteem, social skills, and movement confidence using validated Italian questionnaires. Preliminary data analysis shows promising trends, particularly in proprioception and some PRO scores, specifically in perception of overall health status; this suggests that the sailing intervention may positively influence motor functions, and quality of life. The project is ongoing, and follow-up measurements will help verify long-term effects. This study illustrates the potential of integrating IMUs data and patient-reported outcomes with analytical techniques to assess and monitor the efficacy of non-traditional treatment aimed at improving complex motor impairments in individuals with rare skeletal disorders. As an ongoing proof-of-concept, the study highlights the importance of using quantitative approaches to evaluate the impact of non-standardized intervention methods. The findings could lead the way for the introduction of these multidimensional approaches into clinical practice, contributing to a more diverse and engaging method of rehabilitation.

A novel tool to investigate food compounds and their effects on human health by integrating online bioinformatics resources

Nadia Sanseverino (1,2), Deborah Giordano (1), Angelo Facchiano (1)

(1)National Research Council, Institute of Food Science, 83100 Avellino, Italy. (2) Università di Roma Tor Vergata, Corso di laurea magistrale in Bioinformatica

Numerous studies examine the possible role of food compounds in human health, with particular interest in the prevention of chronic diseases like cardiovascular diseases, neurodegenerative disorders, and cancers. Studies often suggest a preventive effect; however, the underlying biological mechanisms are frequently unclear. The aim of our work was to create a fast, optimized tool for exploring the putative molecular interactions between food compounds and human proteins as potential targets. We integrated various public databases and bioinformatics tools: FooDB for its quality data and API access, GalaxySagittarius-AF, which docks small ligands against domain targets in the curated human proteome database, HProteome-Bsite, while ligand input in SMILES format is handled via PubChem and PUG REST. We developed a user-friendly, cross-platform web interface, built with HTML, PHP, CSS, and JavaScript, enabling easy access for users with varying levels of expertise. Programmatic access to external resources is managed with PHP cURL, while database management is handled by phpMyAdmin connected to an SQL-based database. Registered users can search for foods or compounds. A "Food" search returns a list of related items, which expands to show common and scientific names, food subgroups, and compound lists. Users can select compounds for reverse docking analysis, with results sent via email. A "Compound" search displays compounds with detailed information and reverse docking options, while users can also search human protein domains in HProteome-Bsite by Uniprot ID or protein name. We are expanding the database to include main food sources, 3D structures, biological activities, and predicted ligand-protein interactions. Future plans include integrating additional data from other food composition databases and experimental projects. In this poster, we present the results of a case study, i.e., the search for ericaceous plants and their fruits, such as blueberries.

A bioinformatics strategy to characterize Zostera marina rhythmic genes and explore their conservation between marine and land plants

Alessia Riccardi (1), Luca Ambrosino (2), Marco Miralto (3), Miriam Ruocco (4,5,6), Marlene Jahnke (7), Gabriele Procaccini (1,4), Maria Luisa Chiusano (8), Emanuela Dattolo (1,4)

(1) Department of Integrative Marine Ecology, Stazione Zoologica Anton Dohrn, 80121, Napoli, Italy
(2) Institute of Genetics and Biophysics "Adriano Buzzati-Traverso" (CNR), 8013, Napoli, Italy
(3) Department of Research Infrastructure for Marine Biological Resources, Stazione Zoologica Anton Dohrn, 80121, Napoli, Italy
(4) National Biodiversity Future Centre, Palermo, Italy
(5) Department of Biological, Geological and Environmental Sciences, University of Bologna, Bologna, Italy
(6) Fano Marine Center, Fano, Italy
(7) Department of Marine Science, Tijarnö Marine Laboratory, University of Gothenburg, S-45296 Strömstad, Sweden
(8) Department of Agricultural Sciences, University of Naples Federico II, 80055, Napoli, Italy

As adaptation mechanism to the life on Earth, almost all living organisms synchronize their physiology, behaviour and development with the natural light-dark cycles of ~24 hours. Either the temporal alignment of biological processes to the light-dark cycles and their anticipation are coordinated by an endogenous timekeeper named circadian clock. In plants, light is one of the major input factors involved in the entrainment of the circadian clock. In this complex machinery, the core clock oscillators generate output responses that govern the metabolic activities, the developmental processes and the interaction with biotic/abiotic factors, i.e. stress and photoperiodic responses. While circadian regulation is well-studied in land plants, much less is known in marine ecologically relevant organisms like seagrasses. To fill this gap, we explored the molecular components of the circadian clock in the flowering marine plant Zostera marina using bioinformatic approaches. We produced RNA-seq data from Zostera marina over a 48-hour period under both light-dark (LD) and constant light (LL) conditions. Results were compared to publicly available transcriptional datasets coming from similar experiments conducted under LD and LL in three land plant species: Arabidopsis thaliana, Oryza sativa and Wolffia australiana. To detect 24-hour rhythmic profiles in all the selected species, we used two different tools specifically designed to identify rhythmic patterns in large-scale datasets, i.e. the non-parametric algorithm JTK_CYCLE [1] and the deep learning method BIOCYCLE [2]. Additionally, to infer about the potential evolutionary conservation and divergence existing among the rhythmic genes detected in the 4 selected species, we applied a sequence similarity comparative approach based on an in-house developed tool named COMPARO [3]. Globally, these analyses revealed a lower percentage of circadian-regulated genes in Z. marina compared to the other species, from 27% in Arabidopsis to 2% in Zostera under LL condition. Moreover, only a few core clock oscillators, previously identified in land plants, were conserved in Z. marina. This study provides new insights for the molecular characterization and the evolution of rhythmic genes in marine plants, a topic that remains largely unexplored.

[1] Hughes et al. JTK_CYCLE: an efficient nonparametric algorithm for detecting rhythmic components in genome-scale data sets. J Biol Rhythms. 2010 Oct;25(5):372-80
[2] Agostinelli et al. What time is it? Deep learning approaches for circadian rhythms. Bioinformatics. 2016 Jun 15;32(12):i8-i17
[3] Ambrosino et al. Multilevel comparative bioinformatics to investigate evolutionary relationships and specificities in gene annotations: an example for tomato and grapevine. BMC Bioinformatics 19 (Suppl 15), 435 (2018)

Use of methylation–based machine learning algorithm to classify brain tumors not elsewhere classified (NEC)

Maria Rosaria De Filippo (1), Gianluca Lopez (1,2), Annamaria Morotti (1), Giorgia De Turris (1), Alessandra M. Storaci (1,3), Manuela Caroli (4), Claudia Fanizzi (4), Marco Locatelli (3,4), Stefano Ferrero (1,2), Valentina Vaira (1,3)

(1) Division of Pathology, Fondazione IRCCS Ca' Granda – Ospdeale Maggiore Policlinico, Milan, Italy
(2) Department of Biomedical, Surgical and Dental Sciences, University of Milan, Italy
(3) Department of Pathophysiology and Transplantation, University of Milan, Italy
(4) Division of Neurosurgery, Fondazione IRCCS Ca' Granda – Ospedale Maggiore Policlinico, Milan, Italy

INTRODUCTION DNA methylation involves the addition of a methyl group (CH3) to the DNA. Usually this occurs at the cytosine bases that are followed by a guanine (CpG sites). DNA methylation patterns are often altered in cancer and they can vary among different types or subtypes of tumors. DNA methylation–based machine learning algorithms have been widely described in literature as useful tools to help pathologists classify brain tumors. In particular, the methylation classifier can be used where the tumor diagnosis is more difficult due to ambiguous histology and molecular signatures. These tumors are the so called 'Not Elsewhere Classified (NEC)' brain tumors. Here we use a Random Forest approach to classify our cohort of brain tumors, including three NEC tumors.

METHODS Forty-five primary adult gliomas (IRB#275/2013) were retrieved and analyzed according to WHO CNS 2021 criteria. Methylation profiling was performed using Infinium Methylation EPIC Array and data were analyzed with R v.4.3.1, using a number of packages from Bioconductor and other repositories. Copy number variation analysis was performed using the conumee package from Bioconductor, using as reference baseline a dataset of 50 normal samples downloaded from GEO (GSE157252). To train and validate the Random Forest model we downloaded the entire series of CNS tumors from GEO (GSE109381), where the GSE90496 (2801 samples) was used as training set and the GSE109379 (1104 samples) as validation set. First, we performed background and dye-bias correction. Subsequently, a correction for the type of material tissue (FFPE or frozen) was performed and a number of filters were applied to obtain more cleaned data. The randomForest R package has been used for feature selection starting from filtered data and finally the selected probes (CpG islands) have been used with the validation dataset to assess the accuracy of our classifier. Finally, we use the model to classify our cohort of patients.

RESULTS A total of 10 cases of LGG IDH-wildtype were identified; among those, 3 did not met criteria for GBM and were classified as astrocytoma NEC. According to our classifier, one NEC glioma was labeled as low-grade glioma (LGG-DNT) and two as normal controls (CONTR-Hemi).

CONCLUSION Methylation profiling can support the re-classification of LGG IDH-wildtype if molecular tests are unavailable. Nevertheless, more studies are needed for a more precise clinical categorization of gliomas NEC.

GIMP: A Comprehensive R Package for exploring Genomic Imprinting Methylation Patterns in Imprinting Disorders

Francesco Cecere (1), Abu Saadat (2), Andrea Riccio (1,2), Claudia Angelini (3)

(1) Institute of Genetics and Biophysics (IGB) "Adriano Buzzati-Traverso", Consiglio Nazionale delle Ricerche (CNR), 80131 Naples, Italy

(2) Department of Environmental, Biological and Pharmaceutical Sciences and Technologies (DiSTABiF), Università degli Studi della Campania "Luigi Vanvitelli", Caserta, Italy

(3) Istituto per le Applicazioni del Calcolo (IAC) "Mauro Picone", Consiglio Nazionale delle Ricerche (CNR), 80131 Napoli, Italy

Genomic imprinting is an epigenetic phenomenon where certain genes (imprinted genes, IGs) are expressed differently depending on their parent of origin. The IGs, are controlled by epigenetic marks such as DNA methylation at Imprinting Control Regions (ICRs) differentially established during the gametogenesis in oocytes and sperms. Alterations at imprinted genes cause Imprinting Disorders (IDs), a group of congenital diseases affecting growth, development and metabolism. The DNA methylation analysis methods are crucial for enhancing our understanding of epigenetic regulation in IDs. In particular, methylation arrays produced results comparable to other diagnostic methods, however, there is a lack of effective tools for analysing arrays that focus on ICRs. In response to this need, we introduce GIMP (Genomic Imprinting Methylation Patterns), an R package designed for detailed analysis of ICRs using methylation array data. GIMP offers an automated pipeline for processing and analyzing data from the different methylation arrays platform, including Illumina's 450k, EPIC v1, and EPIC v2 arrays. It would simplify this process from generating CpG sites to the visualization of CpG coverage and identification of DNA methylation abnormalities at ICRs. In a typical workflow, users provide a normalized Betavalue matrix that represents CpG sites across samples; GIMP processes this into ICR-specific analyses. The output includes coverage plots and heatmaps that illustrate differences between experimental groups, aiding in the identification of epigenetic changes linked to the experimental groups. GIMP is adaptable, allowing for custom group assignments and providing extensive visualization and summary tools to enhance result interpretation. GIMP, therefore, represents an integral tool in the investigation of epigenetic regulation of ICRs with its compatibility with standard methylation array platforms. Conclusively, GIMP presents a powerful platform for studying ICRs using methylation array data while bridging an important gap in existing bioinformatics tools.

Predicting medium and long-term risks in asplenic patients: a precision medicine approach

Teresa Cappuccio (1), Maddalena Casale (2), Laura Casalino (3), Maurizio Giordano (1), Marcella Vacca (3), Ilaria Granata (1)

(1) Institute for High-Performance Computing and Networking, National Research Council, Naples, Italy

(2) Haematology and Oncology Pediatric, Department of Women, Children and General and Specialistic Surgery, University of Campania "Luigi Vanvitelli", Naples, Italy

(3) Institute of Genetics and Biophysics "Adriano Buzzati-Traverso", National Research Council, Naples, Italy

The spleen plays a fundamental role in immune response and blood filtration. It recognises and removes malformed, damaged or aged blood cells and responds to pathogens' invasion by producing antibodies and regulating B and T lymphocyte activity. Asplenia, whether surgical, congenital or functional, is associated with several haematological, oncological, immune, and congenital disorders. The loss of splenic function compromises the body's ability to respond to bacterial infections and increases the risk of venous and arterial thrombosis. Severe infections and prothrombotic state due to coagulative and vascular imbalance are relevant and potentially life-threatening complications, with substantial impact on long-term disability as well as public and biological health costs. The Italian National Health Service (SSN) 's annual cost to manage an asplenic patient is considerable due to expenses related to vaccinations, prophylaxis, and potential complications such as infections or thrombosis. Due to the variety of underlying diseases and conditions, predictive factors and tools for associated long-term risks are still lacking. This study aims to develop and validate Machine Learning models for predicting infectious and thrombotic events in asplenic patients. Clinical data from 1,800 patients followed at 42 centers of the "Italian Network for Asplenia (INA)" were collected to this extent. Before training the predictive model, an intensive pre-processing phase of data curation was performed, including transformation, missing data handling, statistical distribution analyses, and association tests. During the model training, feature importance was calculated to improve the model's performance and interpretability. This preliminary study shows promising results in getting new insights for predicting the risk of infectious and thrombotic events in asplenic patients. Implementing predictive models based on relevant clinical features could provide physicians with a useful tool for the preventive management of asplenic patients. This would allow for timely interventions and improve prognosis.

Application of Machine Learning in Predicting Alzheimer's Disease Using Dietary Data

Antonio Agliata (1, 2), Francesco Bardozzo (1), Mariano Caiazzo (2), Gianluca Carotenuto (4), Giovanni Marco Di Vincenzo (1, 2), Angelo Facchiano (3), Antonio Pilato, (2) Mariacarmen Sorrentino (1, 2), Roberto Tagliaferri (1)

1 Dipartimento di Scienze Aziendali, Management and Innovation Systems, Università degli Studi di Salerno, 84084 Fisciano (SA), Italy

2 BC SOFT Centro Direzionale, Via Taddeo da Sessa Isola F10, 80143, Napoli, Italy

3 National Research Council, Institute of Food Science, 83100 Avellino, Italy

4 Università degli Studi di Napoli "Federico II", Napoli, Italy

This work explores the use of machine learning algorithms to predict Alzheimer's disease using dietary data from the National Health and Nutrition Examination Survey (NHANES) for 2011-2012 and 2013-2014. Various supervised learning models, including Random Forest, K-Nearest Neighbors (KNN), and ensemble methods, were employed to classify cognitive performance based on nutrient intake. The dataset, comprising 2994 observations and 38 variables, was pre-processed using techniques such as Multiple Imputation by Chained Equations (MICE) for handling missing data and the Random Over-Sampling Examples (ROSE) algorithm to address class imbalance. Model performance was evaluated using accuracy, sensitivity, and specificity metrics. Results indicate that ensemble methods, particularly Gradient Boosting Machines (GBM), significantly outperformed individual models, achieving an accuracy of 77%. Among the individual models, Random Forest demonstrated the best performance. This study highlights the significant role of dietary models in predicting cognitive decline and demonstrates the effectiveness of advanced machine learning models in medical diagnosis and early intervention strategies for Alzheimer's disease. Index Terms—Machine Learning, Alzheimer's Disease, Dietary Data, NHANES, Supervised Learning, Random Forest, K- Nearest Neighbors, Gradient Boosting, Cognitive Decline, Nutritional Epidemiology.

OMICS techniques to evaluate the toxic effects of biodegradable polymers on two model crustaceans: Idotea balthica basteri Audouin, 1826 and Hippolyte inermis Leach, 1816

Amalia Amato (1,2) , Roberta Esposito (1) , Bruno Pinto (1,3), Thomas Viel (1), Francesca Glaviano (3), Mariacristina Cocca (5), Loredana Manfra (1,6), Giovanni Libralato (2), Emanuele Somma (3), Maurizio Lorenti (4), Eliahu D. Aflalo (7), Amir Sagi (7,8), Maria Costantini (1), Valerio Zupo (3)

(1) Stazione Zoologica Anton Dohrn, Department of Ecosustainable Marine Biotechnology, Via Ammiraglio Ferdinando Acton 55, 80133 Naples, Italy
(2) Department of Biology, University of Naples Federico II, Complesso Universitario di Monte Sant'Angelo, Via Cinthia 21, 80126 Naples, Italy
(3) Stazione Zoologica Anton Dohrn, Department of Ecosustainable Marine Biotechnology, Ischia Marine Centre, 80077 Ischia, Italy
(4) Stazione Zoologica Anton Dohrn, Department of Integrative Marine Ecology, Ischia Marine Centre, 80077 Ischia, Italy
(5) Institute of Polymers, Composites and Biomaterials, National Research Council of Italy, Via Campi Flegrei, 34, 80078, Pozzuoli, Napoli, Italy
(6) Institute for Environmental Protection and Research (ISPRA), Via Vitaliano Brancati 48, 00144, Rome, Italy
(7) Department of Life Sciences, Ben-Gurion University of the Negev, P.O. Box 653, Beer-Sheva 8410501, Israel
(8) Department of Life Sciences, Achva Academic College, Arugot 7980400, Israel

Recent advances in sequencing technology and bioinformatic pipelines offer improved tools to accurately sequence transcriptomes and genomes, opening new perspectives to investigate the responses of marine organisms to environmental xenobiotics. Plastic pollution is a remarkable environmental issue, cause plastic is widespread and often characterized by a long lifetime. Serious environmental problems are caused by the improper management of plastic end-of-life. For this reason, biodegradable polymers (BPs) represent promising materials if correctly applied and managed at their end of life, to reduce environmental problems. However, poor data on the fate and toxicity of BPs on marine organisms still limit their applicability. Aiming at proposing the use of two new model organisms to evaluate the potential toxic effects of BPs in the aquatic environment, we used bioinformatic tools to isolate and identify eighteen genes which are orthologous between two crustaceans, the isopod Idotea balthica basteri Audouin, 1826 (herbivorous scavenger) and the decapod Hippolyte inermis Leach, 1816 (an opportunistic herbivore). This study starts with an analysis of the available information about the genome of I. balthica basteri and the transcriptome of H. inermis. We isolated and identified eighteen genes involved in stress mechanisms and detoxification processes. For each gene, specific pairs of primers were designed, checked by Polymerase Chain Reaction (PCR) and quantified by Real Time qPCR. Previous interactomic analyses indicated that all these genes had several functional interactions, being part of common pathways. Following treatment with five BPs, polybutylene succinate (PBS), polybutylene succinate-co-butylene adipate (PBSA) polycaprolactone (PCL), poly (3-hydroxybutyrates) (PHB) and polylactic acid (PLA), we studied the variation in the expression levels of the genes of interest by Real-Time qPCR. Our results showed that these crustaceans, characterized by different feeding strategies, exhibited different sensitivity to microplastics. More in details, PBSA and PLA had the strongest effects on I. balthica basteri, inducing up-regulation of genes, while PCL and PLA were the strongest for H. inermis, inducing down-regulation of almost all analyzed genes. Noteworthy, this study is the first one adopting omic and molecular approaches to detect crustacean responses to biodegradable polymers. The two species of crustaceans can be considered excellent models and sentinels for the assessment of BPs effects, as well as to provide an early detection of the effects of environmental pollution.

The ongoing MAMELI study: MApping the Methylation of repetitive elements to track the Exposome effects on health

Tiago Nardi (1), Federica Rota (1), Chiara Favero (1), Simona Iodice (1), Luca Pandolfini (2) Elia Biganzoli (1), Valentina Bollati (1)

(1) University of Milan, Milan, Italy
(2) Istituto Italiano di Tecnologia, Genova, Italy

Many studies in epigenetics focus on how DNA modifications, such as methylation, act as response to the "exposome"—i.e. the range of environmental stimuli and stressor conditions—and how these changes contribute to disease development. Previous research has examined how harmful exposures affect the methylation of transposable elements (TEs), often assuming that epigenetic modifications of TEs have inherently pathogenetic effects. In contrast, we hypothesize that the role of TEs as mediators between the exposome and disease is not only harmful. Epigenetic modifications of TEs may actually be part of the physiological/adaptive response to exposures. To test this hypothesis and identify "adaptive TEs," we are conducting a three-phase study involving 6,200 participants from the city of Legnano: - Phase 1. We are currently enrolling 200 participants—healthy blood donors from AVIS (Associazione Volontari Italiani Sangue)—and collecting comprehensive data on their exposome (2 weeks prior biological sample collection). Each participant's TE methylation is being assessed at two time points, spaced six months apart. We are using Nanopore sequencing to identify the differentially methylated TEs. - Phase 2. We will build a statistical model to evaluate the link between the exposome and the methylation of TEs identified in Phase 1, in 2,500 participants. - Phase 3. We will validate the model on further 3,500 participants, identifying lifestyle modifications with the greatest potential for improving health outcomes. Exposome measurements include: - Chemical pollutants (e.g., metals, PFAS) measured via LC-MS/MS in blood and urine samples - Behaviour and habits (e.g., smoking, diet) assessed through questionnaires and monitored using a personal app - Physiological parameters (e.g., heart rate variability, sleep quality) and physical activity recorded with a wearable device - Air pollution exposure, estimated using participants' GPS data and integrating regional pollution estimates with data from monitoring stations deployed in Legnano for the study The MAMELI project sets unique opportunities on data integration and analysis according to the above hypotheses. To analyze the multilevel/multiomic data of the study we will deploy various strategies, including methods based on network modelling and multivariate analysis of sparse data in presence of non-linear and non-additive effects.

circQIT: circular RNA Quantification and Identification Tool

Domenico Palumbo (1), Francesca Calanca (1), Viola Melone (1), Luigi Palo (1,2), Alessandro Giordano (1,2), Dilia Rea (1), Francesca Rizzo (1), Giovanni Nassa (1), Alessandro Weisz (1,2,3), Roberta Tarallo (1)

(1) Laboratory of Molecular Medicine and Genomics, Department of Medicine, Surgery and Dentistry "Scuola Medica Salernitana", University of Salerno, 84081 Baronissi (SA), Italy
(2) Genome Research Center for Health (CRGS), 84081 Baronissi (SA), Italy
(3) Medical Genomics Program and Division of Oncology, AOU "S. Giovanni di Dio e Ruggi d'Aragona", Università di Salerno, 84131 Salerno, and Rete Oncologica Campana, Italy

Circular RNAs (circRNAs) are covalently closed RNA molecules that regulate gene and protein expression both at the transcriptional and post-transcriptional levels, and act as sponges for microRNAs. Despite their considerable potential, research on circRNAs remains limited. However, increasing insights into their role in disease pathogenesis have heightened interest in their identification and functional analysis through bioinformatics methods. Identifying circRNAs involves searching for back-splicing sites, with each detection tool employing a unique combination of strategies. This leads to analytic variability affecting sensitivity, precision, and computational cost. Indeed, to enhance results' reliability, it is advisable to apply a combination of multiple algorithms. However, this means that the user needs to install and test several tools written in different codes, which often results in a waste of time. In our pipeline, we have integrated several tools, within a Docker container, that are useful not only for circRNA identification but also for their quantification and annotation, the last two topics being always challenging due to a lack of scientific consensus. Additionally, we have also proposed a subsequent step for functional enrichment and, moreover, the circRNAs identified in common by the different tools can be also represented to the user in tables and graphs. CircQIT performance has been tested both on publicly available data and on in-house generated RNA-seq data with encouraging results.

Non-invasive approaches for 3D Reconstruction of Tree Root Systems in Urban Environments

Rosangela Casolare, Gabriella Sferra, Stefano Ricciardi, Rocco Oliveto, Gabriella Stefania Scippa

University of Molise, Italy

Trees play a vital role in promoting biodiversity within urban environments by providing habitat, food, and shelter to various animal species, while also functioning as essential “green infrastructure” that delivers a range of ecosystem services. The establishment, monitoring, and management of biodiversity centered around trees are crucial for ensuring healthy and sustainable urban ecosystems. The health and stability of trees are closely linked to the condition of their root systems. Therefore, analyzing the morphology and physiology of tree roots is fundamental for maintaining ecological balance. However, in urban settings, trees face a broader range of abiotic and biotic stresses compared to natural environments. Root systems in cities often encounter poor-quality soils, limited soil volume, and physical barriers, making their health particularly vulnerable. Currently, the most common method for studying root systems involves excavation and soil removal, which is invasive and may cause harm. As such, the need for non-invasive techniques is becoming increasingly urgent. In geophysical sciences, powerful, close-range remote sensing technologies are already being used to detect subsurface objects, such as roots and pipes, by analyzing changes in physical properties. Ground-penetrating radar (GPR) and electrical resistivity tomography (ERT) are two non-destructive methods that have shown significant potential for imaging and reconstructing buried structures in three dimensions. Nevertheless, the accuracy and efficacy of these techniques in relation to root morphology and architecture, are often hindered by the lack of standardized protocols for data analysis and interpretation. In this context, accurate 3D reconstruction, based on 2D sections from radargrams (graphical representations of the acquired signals), is essential for correctly interpreting the data and identifying root patterns. As part of the Italian Recovery and Resilience National Plan (RRNP), spearheaded by the National Biodiversity Future Centre (NBFC) of which the University of Molise is partner, we aim to develop a non-invasive automated tool to reconstruct 3D models of tree root systems for monitoring urban biodiversity. At this scope we will evaluate and compare the models resulting from GPR and ERT signals, with those obtained through laser scanning techniques post-excavation, ensuring accurate representation and enhanced understanding of urban root system dynamics.

The Probabilistic Patient Navigator: An Interpretable Approach to Healthcare Decision-Making

Sheresh Zahoor (1), Pietro Liò (2), Gaël Dias (3), Mohammed Hasanuzzaman (4)

(1) Munster Technological University, Cork, Ireland
(2) University of Cambridge, Cambridge, UK
(3) University of Caen Normandy, Caen, France
(4) Queens University Belfast, Northern Ireland, UK

Healthcare decision-making is of paramount importance, as decisions profoundly impact patient well-being. In this context, interpretable and transparent AI models play a crucial role, fostering trust among both clinicians and patients. This work introduces the Probabilistic Causal Fusion (PCF), a novel interpretable AI approach leveraging the synergy between Causal Bayesian Networks (CBNs) and ensembles of Probability Trees (PTrees). PCF harnesses the causal order learned by the CBN to structure PTrees, capturing inherent causal relationships and enabling domain knowledge incorporation through counterfactuals. Validated on three real-world healthcare case studies (MIMIC-IV, Framingham Heart Study, Diabetes), PCF achieves comparable prediction performance to benchmarks. These case studies were selected for their comprehensive and diverse clinical data, encompassing a wide range of variables critical for evaluating the effectiveness of PCF in different medical contexts. The true value of PCF lies beyond mere prediction. Unlike traditional models, PCF empowers clinicians with a comprehensive toolkit for exploring hypothetical interventions and counterfactual scenarios. This enhanced understanding of risk factors, interventions, and variable interactions significantly enhances clinical decision-making. To deepen interpretability, our framework integrates sensitivity analysis and SHapley Additive exPlanations (SHAP). Sensitivity analysis within the CBN component assesses the influence of parameters on target variables, providing insights into factors affecting Length of Stay (LOS), Coronary Heart Disease (CHD), and Diabetes. SHAP values, applied to the predictive models, offer a unified measure of feature importance for each prediction, enabling transparent and actionable insights into the decision-making process. By surpassing the limitations of interpretable AI development in healthcare, PCF fosters generalisability across diverse settings. This transformative potential holds significant promise for revolutionising healthcare delivery and ultimately improving patient outcomes.

Exploring biomarker dynamics in Inflamm-Aging phenomenon: A longitudinal multivariate analysis from the HEBE project

Davide Biganzoli, Simona Iodice, Federica Rota, Valentina Artusa, Lara De Luca, Claudio Fenizia, Daniela Lucini, Chiara Mandò, Francesca Bianchi, Mario Clerici, Valentina Bollati, Elia Biganzoli, HEBE Consortium

University of Milan

HEBE (Healthy aging versus inflamm-aging: the role of physical Exercise in modulating the Biomarkers of age-associated and Environmentally determined chronic diseases) project addresses the phenomenon of inflamm-aging, a chronic low-grade inflammation process that accelerates cellular decline and contributes to the development of age-related diseases1. In this longitudinal study, we analyzed data from 100 participants at two time points (T0, pre-intervention, and T1, post-intervention), following the implementation of a personalized physical exercise program, to evaluate the dynamic changes in cytokines and other key biomarkers. Using multivariate statistical methods, including Principal Component Analysis (PCA) and its extensions, along with Repeated Measures Analysis of Variance-Simultaneous Component Analysis (RM-ASCA), we assessed global patterns and interrelationships among inflammatory and non-inflammatory markers, such as cytokines, telomere length, and the oxidative stress marker 8-OHdG. These analyses enabled us to capture the complex biomarker interactions driving inflamm-aging and to identify potential modifiable factors. Our findings suggest that physical exercise, coupled with lifestyle changes, significantly modulates key biomarkers, promoting resilience against age-related inflammatory processes and reducing the burden of non-communicable diseases (NCDs)2. The application of multivariate techniques was essential not only for modeling the temporal trends of these biomarkers but also for providing a more comprehensive and integrated analysis of their interrelationships. This approach revealed latent patterns and clustering effects within the data that would have remained hidden using traditional univariate methods, thereby offering deeper insights into the complex interactions among the biomarkers. This integrative analysis underscores the relevance of inflammation control in aging and highlights physical exercise as a critical intervention to mitigate its detrimental effects. These findings offer novel insights into the systemic responses triggered by exercise and present actionable strategies for preventing age-related decline through targeted lifestyle interventions.

A computational approach to the dissection of cardiopharyngeal mesoderm differentiation involved in congenital heart disease

Olga Lanzetta (1), Marchesa Bilio (1), Johannes Liebig (2) , Katharina Jechow (2), Foo Wei Ten (2), Rosa Ferrentino (1), Ilaria Aurigemma (3), Elizabeth Illingworth (3), Christian Conrad (2), Sören Lukassen (2), Antonio Baldini (5), Claudia Angelini (4)

1) Università di Genova, Italy
2) Berlin Institute of Health at Charité – Universitätsmedizin Berlin, Center of Digital Health, Berlin, Germany
3) Dept. of Chemistry and Biology, University of Salerno, Italy
4) Istituto per le Applicazioni del Calcolo, CNR, Naples, Italy
5) Dept. of Molecular Medicine and Medical Biotechnology, Univ. Federico II, Naples, Italy

Background TBX1 is a T-box transcription factor required to develop the pharyngeal apparatus. The Tbx1 gene acts as a primary marker for the cardiopharyngeal mesoderm (CPM). This multipotent cell population is responsible for generating progenitors of the second heart field and branchiomeric muscles. Despite TBX1's importance, the molecular mechanisms that affect CPM differentiation in mammals remain largely unclear. Methods To explore the role of TBX1 in CPM differentiation, we conducted experiments using wild-type and Tbx1-/- mouse embryonic stem cells (ESCs) differentiated in culture. We collected multiomic data at two critical stages of differentiation, employing simultaneous single-cell RNA sequencing (scRNA-seq) and ATAC sequencing (scATAC-seq) to capture changes in transcriptional activity and chromatin landscapes. We integrated these datasets using the Seurat pipeline and constructed gene cluster-specific co-expression networks using the hdWGCNA package. We then analyzed key hub genes from these networks to evaluate chromatin responses based on genotype. We set up a machine-learning model that identified putative enhancers linked to the hub genes, and enriched transcription factor motifs were determined using Homer. RNA velocity analysis implemented by Dynamo was utilized to infer the directionality of cell state transitions and transcriptional dynamics, while Monocle3 was used to reconstruct cellular trajectories. Results and Conclusions Our analysis demonstrated that TBX1 regulates chromatin accessibility and gene expression within an evolutionarily conserved transcriptional module crucial for trunk and pharynx development. This module includes genes such as Tea Shirt (Tshz), Sine Oculis (Six), Eye Absent (Eya), and Ebf/Collier factors. Using a machine learning approach, we found evidence that TBX1 modulates this transcriptional network through SIX factors. Although the absence of TBX1 did not significantly alter the overall cellular trajectories, it disrupted the maintenance of appropriate transcriptional profiles, activating ectopic epithelial/ectodermal and extracellular matrix genes in a subset of mesodermal progenitor cells. Overall, the multi-omic analysis, based on different tools and techniques, allowed us to examine the data from multiple perspectives, revealing key insights into the regulatory mechanisms and the influence of TBX1 in CPM differentiation. Our findings highlight the essential role of TBX1 in sustaining coherent cell fate trajectories and preventing transcriptional drift during CPM differentiation, significantly impacting chromatin structure and gene regulation.

Using Machine Learning to Explore the Genetics of Tobacco Addiction: A Study of an African American Population

Ibrahima Diallo (1,2), Ines Abdeljaoued-Tej (1), Harouna Soumaré (1), Amor Mosbah (3), Mariem Hanachi (2), Oussema Souiai (2), Emna Harigua (4), Christian Happi (1), Alia Benkahla

(1) African Center of Excellence for Genomics of Infectious Diseases, Redeemer's University
(2) Laboratory of BioInformatique, bioMathematique and bioStatistique, Pasteur Institute of Tunis
(3) Laboratory of Biotechnology and Bio-Geo Resources Valorization, Biotechpol of Sidi Thabet
(4) Laboratory of Molecular Epidemiology and Experimental Pathology, Pasteur Institute of Tunis

The World Health Organization estimates that lung cancer from tobacco smoking causes more than 7 million deaths each year worldwide. Cigarette smoking is also a risk factor for respiratory tract and other infections, osteoporosis, reproductive disorders, adverse postoperative events and delayed wound healing, duodenal and gastric ulcers, and diabetes. Most tobacco smokers would like to stop smoking. So trying to understand the genetic phenomena underlying this addiction would be a good initiative for finding a therapy to fight tobacco addiction. This poses challenges in accurately predicting the phenotype solely based on individual susceptibility variants. With this perspective in mind, studies have been carried out to prove that smoking has a heritability rate ranging from 21% to 84%. So it must be possible to identify the genetic factors associated with tobacco addiction. There are various approaches to achieving this like genome-wide linkage analysis, candidate gene-based association, and genome-wide association studies (GWAS). The application of machine learning (ML) in the field of bioinformatics was investigated, with a special focus on the development of classification models to distinguish smokers from non-smokers based on their genetic profiles. A dataset from the GEO database was used. It included genetic information and smoking status of African American participants. The data were pre-processed, including cleaning and imputing missing values. Subsequently, feature selection techniques were used to identify the most relevant variables for inclusion in our artificial intelligence algorithms. Various ML methods, such as support vector machines (SVM) and logistic regression, were used to train classification models on the given training data set. In this context, our study focuses on single nucleotide polymorphism (SNP)-based phenotype profiling of individuals of African origin. To achieve this goal, we used a variety of ML tools, such as Elastic Net (EN) for feature selection and Logistic Regression (LR) for smoking status classification based on SNPs data. Finally, we identified the genes associated with the SNPs found to perform functional annotation and check whether the genes obtained are associated with nicotine dependence. Results showed robust classification performance, especially in logistic regression and SVM models, effectively discriminating smokers from nonsmokers. This model is expected to have a significant impact on smoking prevention. In addition, future research would benefit from examining the potential generalizability of this model to diverse ethnic populations.

Optimizing dual RNA-Seq for phylogenetically similar organisms: a comparative analysis of sequential and combined mapping in host-parasitic plant interactions

Gaetano Aufiero (1), Carmine Fruggiero (1,2), Davide D'Angelo (1), Edoardo Pasolli (1), Nunzio D'Agostino (1)

(1) Department of Agricultural Sciences, University of Naples Federico II, Piazza Carlo di Borbone 1, 80055 Portici, Napoli, Italy
(2) Department of Electrical Engineering and Information Technology, University of Naples Federico II, Via Claudio 21, 80125 Napoli, Italy

Transcriptome studies on host-parasitic plant interactions have traditionally involved analyzing the two organisms separately, often using laser capture microdissection (LCM) to isolate specific tissues from host and parasite. While effective, LCM is expensive, time-consuming, and requires specialized skills. Dual RNA-seq presents a promising alternative, enabling the in silico separation of mixed transcripts from the interface region without the need for physical dissection of infected tissues. This study explores the feasibility of dual RNA-seq in plant-plant parasitic systems by simulating interactions between Arabidopsis thaliana - Cuscuta campestris and Solanum lycopersicum - C. campestris. The research addresses challenges such as multiple mapping and cross-mapping of reads between host and parasite genomes, particularly as the evolutionary divergence between interacting organisms decreases. Two mapping approaches, sequential and combined, were evaluated using artificial datasets and assembled reference genomes. Both methods achieved high mapping rates (~90%) and low cross-mapping rates (~1%), with the combined approach offering advantages in processing time and minimizing cross-mapping errors. The study found that the evolutionary distance between parasitic and host plants did not significantly impact the accuracy of read assignment to their respective genomes, as there were enough polymorphisms to allow for reliable differentiation. These findings confirm that dual RNA-seq is a reliable and effective method for studying interactions between organisms, even when both organisms belong to the same taxonomic kingdom. This sets the foundation for further research into the key genes involved in plant parasitism.

Structural Basis of PMM1 and PMM2 Dimerization Specificity: Implications for Glycosylation Disorders

Jessica Bovenzi (1), Maria Monticelli (1,2), Giuseppina Andreotti (2), Maria Vittoria Cubellis (1), Bruno Hay Mele (1)

(1) Biology Dept. ; Università degli Studi di Napoli Federico II; Complesso Universitario MSA, Ed. 7; Via Cinthia, 26 80126 Naples (NA)

(2) Institute of Biomolecular Chemistry ICB, CNR, Via Campi Flegrei 34, 80078, Pozzuoli, Italy

Glycosylation plays a fundamental role in cellular function by affecting the stability, transport, and activity of proteins. Phosphomannomutases (PMM) are responsible for the conversion of mannose-6-phosphate to mannose-1-phosphate. Humans have two paralog phosphomannomutases, PMM1 and PMM2; although they may perform overlapping functions in vitro, it is accepted that the physiological role of PMM1 in vivo would be that of bisphosphatase, rather than phosphomannomutase. Currently, no direct pathological conditions have been associated with mutations in PMM1, while mutations in PMM2 are pathologically relevant, leading to the PMM2 associated congenital disorder of glycosylation (PMM2-CDG), and the presence of wt-PMM1 does not compensate for the lack of PMM2. The defective glycosylation disrupts proper protein function, impacting tissue and organ development. PMM2-CDG patients are typically composite heterozygous, and it has previously been demonstrated that dimers can be made of subunits carrying different mutations. Interestingly, some kind of PMM2 heterodimers have higher specific activity than the related homodimers, thus the structural study of the dimerization surface is highly interesting. To unravel the regulatory and functional mechanisms of PMM1 and PMM2, it is interesting to investigate whether PMM1 and PMM2 can heterodimerize under physiological conditions, and which are the factors eventually involved in their dimerization specificity. To evaluate whether differences in the residues at the dimerization interface enable the recognition of identical subunits while limiting interactions with different subunits, we used AlphaFold3 to build models of homodimers of PMM1 and PMM2 and of a PMM1-PMM2 heterodimer. The analysis of those structures with EMBL PISA tool gave insights on dimer stability and permitted comparison of the different combinations. Data were compared with previously published experimental results, that had showed a stronger thermal stability of wt-PMM2 homodimer with respect to wt-PMM1 homodimer. The predicted stability values were comparable across the three combinations, suggesting that heterodimerization is possible when the two PMMs coexist. The analysis revealed that the Ala94 residue of PMM2 plays a unique role in heterodimer formation but is not involved in the homodimerization of PMM2. However, no significant differences were found in the interfacing residues to suggest that these residues are responsible for dimerization selectivity. This raises further questions about why PMM1 and PMM2 predominantly form homodimers rather than heterodimers.

Clustering-Based Stratification of Mild Cognitive Impairment: Insights from Blood Transcriptomic Data

Laura Antonelli, Claudia Di Napoli, Lucia Maddalena, Giovanni Paragliola, Patrizia Ribino, Luca Serino, Ilaria Granata

Ist. di Calcolo e Reti ad Alte Prestazioni (ICAR) del CNR

Alzheimer's disease (AD) is a complex brain disorder that causes progressive cognitive decline and neuropsychiatric complications, ranking among the top 10 causes of death. Clinically, there are three stages of cognitive decline: dementia, mild cognitive impairment (MCI), and subjective cognitive decline (self-reported cognitive issues). Both MCI and subjective cognitive decline are recognized as risk factors for dementia, but most countries do not recommend specific treatments outside of clinical trials. Identifying preventive indicators of AD is critical for developing effective disease control strategies. The diagnosis of MCI or dementia is typically made by a clinician through detailed neuropsychological and neurological assessments, basic lab tests, and structural brain imaging (MRI or CT scans). Additional diagnostic methods may include biomarkers from PET brain scans or cerebrospinal fluid (CSF). Since CSF collection requires lumbar punctures, finding less invasive blood-based biomarkers is crucial. A key goal of precision medicine, though highly challenging, is to improve the classification of pathological conditions based on molecular markers rather than relying solely on clinical symptoms. From this perspective, providing a clear and objective identification of conditions that could progress into more severe symptoms while distinguishing them from other types of impairments is crucial for improving early detection of AD. A clustering-based molecular stratification of MCI patients by analyzing transcriptomic data from blood samples provided by the ADNI (Alzheimer's Disease Neuroimaging Initiative) resource is realized with this aim. A specific AI-based pipeline is built to perform the transcriptomic analysis, integrating unsupervised and supervised learning approaches. Starting with feature selection techniques, the pipeline uses deep learning-based autoencoders to embed gene expression data from 381 patients into a lower-dimensional latent space. This latent representation is then fed into a clustering algorithm to determine the final result. The identified MCI patient clusters show significant molecular, clinical, and pathophysiological differences. Differences are also evident between some MCI patient clusters compared to patients with AD diagnosis, as well as with control samples. To our knowledge, this is the first study to demonstrate a molecular-based stratification of MCI patients using blood transcriptomic data. These findings can lead to an early and precise diagnosis of dementia and related conditions, offering a more objective alternative to standard cognitive-based definitions often influenced by subjective interpretation.

Structure-based design of peptides for biomedical applications

Eleonora Proia (1), Emanuele Savino (2), Andrea Di Giulio (2), Gianmarco Pascarella (1), Andrea Ilari (1), Allegra Via (2), Veronica Morea (1), Patrizio Di Micco (1)

(1) Institute of Molecular Biology and Pathology (IBPM) of the National Research Council of Italy (CNR)
(2) Department of Biomedical Sciences "A. Rossi Fanelli", Sapienza University of Rome, Italy

Peptides are highly effective research tools and attractive therapeutic agents per se, as well as convenient lead compounds for the rational development of non-peptidic small molecule drugs. Indeed, since interaction surfaces are selected by evolution to be highly specific for specific interacting molecules and, in many cases, to bind these interactors with high affinity, peptides mapping on protein regions involved in interactions with macromolecular partners are usually effective interaction mimics and/or inhibitors. For these reasons, when a three-dimensional (3D) structure of a protein complex is known, peptides can be designed to target interaction sites, preventing solubility issues through empirical rules. Even in cases where only the free state of a protein structure is known, peptides aimed at inhibiting protein interactions and/or binding the interaction partners of the protein, can be designed to cover the entire protein surface and tested experimentally. In the past years, our structure-based design procedure allowed us to obtain peptides endowed with several biomedical activities.

1) We designed a 16-residue peptide based on the crystal structure of leucyl-tRNA synthetase (LeuRS) and tRNALeu from Thermus thermophilus, which binds to human mitochondrial tRNALeu(UUR) and other mutant tRNAs linked to severe mitochondrial diseases like MELAS and MERRF. This peptide helps the mutated tRNAs regain a native-like conformation, restoring their function and rescuing the pathological phenotype in cellular models at 0.1 μM concentration. A peptide-mimetic derived from this peptide has been patented and is under investigation by a dedicated company.

2) Twelve peptides were designed to encompass the solvent-accessible surface of VEGFR-1's second immunoglobulin-like domain, utilizing its 3D structure with VEGF-A. One peptide exhibited strong pro-angiogenic activity through α5β1 integrin interaction, while two others demonstrated anti-angiogenic effects by inhibiting sVEGFR-1 dimerization in one case, and its interaction with NRP-1, in the other. These peptides show potential in modulating angiogenesis at low μM concentrations.

3) More recently, we designed peptides mapping on the surface of the receptor binding domain of the SARS-CoV-2 spike protein, whose 3D structure has been determined in multiple experiments, which have been used to identify the epitope of neutralizing antibodies.

To increase the speed of our structure-based peptide design procedure and make it available to the scientific community, we developed a software that can be used in two modes. The automated mode allows users to select a protein structure to generate a comprehensive set of peptide mappings. In the interactive mode, users can customize parameters like peptide length, overlap, positioning relative to secondary structures, and criteria for interaction identification. A public server for this software is available at:
https://minepept.ibpm.cnr.it/.

Omics tools to isolate genes involved in sex differentiation of the shrimp Hippolyte inermis (Leach, 1815): a new challenging approach

Bruno Pinto (1,2), Marialuisa Lusito (1), Amalia Amato (1,3), Giuseppe Trotta (4,5), Roberta Esposito (1), Elvira Brunelli (5), Amir Sagi (6), Eliahu D. Aflalo (6,7), Takashi Gojobori (8), Robert Hoehndorf (8), Maria Costantini (1), Valerio Zupo (9)

1) Stazione Zoologica Anton Dohrn, Department of Ecosustainable Marine Biotechnology, Via Ammiraglio Ferdinando Acton 55, 80133 Naples, Italy
2) Stazione Zoologica Anton Dohrn, Department of Ecosustainable Marine Biotechnology, Ischia Marine Centre, 80077 Ischia, Italy
3) Department of Biology, University of Naples Federico II, Complesso Universitario di Monte Sant'Angelo, Via Cinthia 21, 80126 Naples, Italy
4) Department of Research Infrastructures for Marine Biological Resources, Stazione Zoologica Anton Dohrn, Calabria Marine Centre, C. da Torre Spaccata, 87071, Amendolara, Italy
5) Department of Biology, Ecology and Earth Sciences, University of Calabria, 87036, Arcavacata di Rende (CS), Italy
6) Department of Life Sciences, Ben-Gurion University of the Negev, P.O. Box 653, Beer-Sheva 8410501, Israel
7) Department of Life Sciences, Achva Academic College, Arugot 7980400
8) Computational Bioscience Research Center (CBRC), King Abdullah University of Science and Technology, Thuwal, Saudi Arabia,
9) Stazione Zoologica Anton Dohrn, Department of Ecosustainable Marine Biotechnology, Ischia Marine Centre, 80077 Ischia, Italy

The sexual differentiation in the protandric hermaphrodite Hippolyte inermis (Leach, 1815) became a hot topic in the last decades. This decapod has been extensively used to elucidate the effect of algal compounds on the physiology of crustaceans. It has been also studied to improve our theories on the ecological importance of the proterandric sex reversal. The ingestion of specific benthic diatoms (Cocconeis spp.) induces the early destruction of the shrimp's androgenic gland (AG) due to programmed cell death (PCD). In particular ferroptosis, followed by apoptosis, is the main PCD process leading to sex reversal of these shrimps, upon the ingestion of selected diatoms. Male and female phenotypes of shrimps often show different sizes because males are often smaller. The presence of masculine appendix on the second pleopod of adult males is an external sexual character, but its identification requires the dissection of mature individuals. Sex-specific genes related to Insuline like hormone (IAG) and Vitellogenin (VG), together with an Uncharacterized male gene (UCM) and a female gene (UCF), were recently isolated, starting from a wide transcriptomic analysis. Here, for the first time we isolated other two uncharacterized sex-specific genes isolated from eyestalks, UCMe and UCFe. We set an effective extraction method to obtain sufficient amounts of high?quality RNA. Further, specific primers were designed for housekeeping genes and the two genes of interest. The fragments were amplified by Polymerase Chain Reaction (PCR) and purified in agarose gel, and their specificity was checked by DNA sequencing. The variation in expression of all the genes under analysis was checked by Real Time qPCR on male and female shrimps. The vitellogenin gene was the most valid to be used as a molecular tool for the quick determination of sex in adult shrimp, overcoming the long-lasting and difficult identification of sexual appendices through microscope observations. The expression levels of the same gene could be quantified in future in pools of individuals to obtain a quick evaluation of the sex ratios in natural populations.

The ALFI database: images and annotations for label-free time-lapse microscopy

Federica Polverino (1), Laura Antonelli (2), Alexandra Albu (3), Aroj Hada (3), Italia A. Asteriti (1), Francesca Degrassi (1), Lucia Maddalena (2), Mario R. Guarracino (3), Giulia Guarguaglini (1)

1. IBPM, National Research Council, Italy
2. ICAR, National Research Council, Italy; 3. University of Cassino and Southern Lazio, Italy

Microscopy represents a high-content methodology with applications at multiple levels in biological research, thanks to recent and unprecedented developments that have amplified its informative power. The drug discovery and cancer research fields have largely benefited from innovative image-based approaches. Time-lapse imaging of living cells is used to observe the dynamic behavior of single cells over time, providing a key contribution to the understanding of the heterogeneity of cellular responses to specific treatments. In these analyses, the challenge of detecting and tracking multiple moving objects in a video is made even more arduous as cells change their morphology over time, can partially overlap, and mitosis leads to new cells. Furthermore, the application of such analyses to extended data sets requires automated approaches based on machine learning/artificial intelligence to define complex phenotypic profiles. Available automated systems for detecting cellular events such as cell division and cell death mostly rely on the use of fluorescently labeled cells. This limits the flexibility of applications due to phototoxicity effects, cell line availability, and stability of fluorescent probes. In this study, we present the applications and use of ALFI, a dataset of images and annotations for label-free microscopy, that we recently made publicly available to the scientific community. ALFI notably extends the current panorama of expertly labeled data for the detection and tracking of cultured living cells, consisting of time-lapse image sequences, acquired by differential interference contrast microscopy, of nontransformed and cancer human cells under different experimental conditions. It contains various annotations for the identification of interphase and mitotic cells, mitotic defects, and cell death (pixel-wise segmentation masks, object-wise bounding boxes, and tracking information). The dataset is useful for testing and comparing methods for identifying interphase and mitotic events and reconstructing their lineage, and for discriminating different cellular phenotypes using label-free techniques.

Pipeline for RNA-seq data analysis on organoids derived from celiac patients

Davide Biondi (1), Simone Bonora (2), Francesco Bardozzo (1), Anna Marabotti (2), Roberto Tagliaferri (1)

1: Dipartimento di Scienze Aziendali - Management and Innovation Systems, Università di Salerno, Fisciano (SA), Italy
2: Dipartimento di Chimica e Biologia "A. Zambelli", Università di Salerno, Fisciano (SA), Italy

Celiac disease is an autoimmune disorder triggered by gluten in genetically predisposed individuals. The aim of our work was to develop a new RNA-seq analysis pipeline, integrating different tools and derive new insights from an existing dataset, publicly available in the GEO NCBI database (GEO accession: GSE113492), in which RNA sequencing on organoids derived from duodenal biopsies of celiac (CD) and non-celiac (NC) patients was performed and a differential expression analysis (DEA) was obtained using EdgeR, an R/Bioconductor software package for differential analyses of sequencing data in the form of read counts for genes or genomic features. The preprocessing of RNA-seq data involved adapter trimming, quality control using Trim Galore, and removal of rRNA sequences with SortMeRNA. Reads were mapped to the human genome using STAR, and gene expression was quantified with HTSeq. DEA was performed using pyDESeq2, a python implementation of the DESeq2 method for DEA with bulk RNA-seq data, originally in R; Gene Set Enrichment Analysis (GSEA) was performed through GSEApy, and functional annotation via DAVID was applied to interpret the biological significance of these genes. The analysis revealed significant differences in the gene expression profiles between CD and NC organoids, with 83 genes being differentially expressed. Key findings include the downregulation of genes like DKK1, a WNT negative regulator, and Trefoil Factors (TFF1, TFF2), which are crucial for maintaining mucosal integrity. Conversely, genes such as NLRP6, involved in the inflammasome complex, and DEFA6 (Human alpha defensin 6), which has a role in bacterial defense, were found to be overexpressed in CD organoids. This highlights the heightened inflammatory state in CD even in the absence of gluten, a characteristic that may contribute to the disease's pathology. Additionally, genes linked to immune responses, such as HLA-DRB1, which is associated with CD-related HLA haplotypes, were significantly overexpressed, underscoring the genetic predisposition of CD patients. Literature confirmed the association between HLA-DRB1 alleles and both celiac and rheumatoid arthritis (RA) diseases, and considering the high prevalence of CD in RA patients compared to the healthy population, a possible shared autoimmune mechanism between the diseases could involve HLA-DRB1. The study also explored how the overexpression of CD36, a receptor involved in fatty acid metabolism, and its interaction with thrombospondin-1 could influence angiogenesis and nutrient absorption in CD patients, contributing to common symptoms like malnutrition. These findings partly confirmed the results of the original study, while also identifying new differentially expressed genes that are consistent with the celiac disease condition. Furthermore, our results were largely validated using the GEO2R tool available on the GEO NCBI platform that uses DESeq2 to perform DEA, reinforcing the reliability of our analysis.

PTF-Vāc: Ab-initio discovery of plant transcription factors binding sites using explainable and generative deep co-learning encoders-decoders

Sagar Gupta (1,2), Jyoti (1,2), Umesh Bhati (1,2), Veerbhan Kesarwani (1,2), Akanksha Sharma (1), Ravi Shankar* (1,2)

1 Studio of Computational Biology & Bioinformatics, The Himalayan Centre for High-throughput Computational Biology, (HiCHiCoB, A BIC supported by DBT, India), Biotechnology Division, CSIR-Institute of Himalayan Bioresource Technology (CSIR-IHBT), Palampur (HP), 176061, India.
2 Academy of Scientific and Innovative Research (AcSIR), Ghaziabad, Uttar Pradesh- 201002

Discovery of transcription factors (TFs) binding sites (TFBS) and their motifs in plants pose significant challenges due to high cross-species variability. The interaction between TFs and their binding sites is highly specific and context dependent. Most of the existing TFBS finding tools are not accurate enough to discover these binding sites in plants. They fail to capture the cross-species variability, interdependence between TF structure and its TFBS, and context specificity of binding. Since they are coupled to predefined TF specific model/matrix, they are highly vulnerable towards the volume and quality of data provided to build the motifs. All these software make a presumption that the user input would be specific to any particular TF which renders them of very limited uses. This all makes them hardly of any use for purposes like genomic annotations of newly sequenced species. Here, we report an explainable Deep Encoders-Decoders generative system, PTF-Vāc, founded on a universal model of deep co-learning on variability in binding sites and TF structure, PTFSpot, making it completely free from the bottlenecks mentioned above. It has successfully decoupled the process of TFBS discovery from the prior step of motif finding and requirement of TF specific motif models. Due to the universal model for TF:DNA interactions as its guide, it can discover the binding motifs in total independence from data volume, species and TF specific models. PTF-Vāc can accurately detect even the binding motifs for never seen before TF families and species, and can be used to define credible motifs from its TFBS report.

* A copy of the same abstract with full text has also been made available at BioRxiv (https://www.biorxiv.org/content/10.1101/2024.01.28.577608).

A Marine Automated Recognition and Identification System based on Multimodal Artificial Intelligence Methods

Simone Cioffi, Federica Ferrigno, Gennaro Mellone, Vincenzo Mariano Scarrica, Emanuel Di Nardo, Roberto Sandulli, Antonino Staiano, Angelo Ciaramella

University of Naples "Parthenope", Italy

The need for autonomous software and algorithms for ecosystem observations has increased significantly with the advancement of neural pattern recognition and machine learning technologies. In this work, the Marine Automated Recognition and Identification System (MARIS) is designed to address this need by providing a comprehensive and optimized platform for the detailed study and monitoring of marine flora and fauna. Leveraging state-of-the-art neural network and computer vision techniques, MARIS aims to provide accurate and real-time identification of marine species, track biodiversity, and assess the health of marine ecosystems. Beyond ecological monitoring, MARIS also includes capabilities to detect and classify marine debris, such as plastics and other pollutants, contributing to pollution control and environmental protection efforts. The system integrates custom Convolutional Neural Networks and Vision-Language Models (VLMs), to process complex visual data with remarkable accuracy. \\ The system's object detection component effectively identifies and localizes diverse marine species and other underwater objects, with high precision. The fine-tuning processes have demonstrated MARIS' ability to adapt to new data and improve its detection and classification capabilities over time. In addition, the Visual Question Answering (VQA) feature of MARIS demonstrates the model's ability to interpret and respond to user queries about marine imagery, providing contextually relevant and accurate information. This capability is critical for facilitating rapid analysis and decision-making in marine research, contributing to the overall utility of MARIS as a tool for scientists and conservationists. In conclusion, MARIS is proving to be a powerful tool for the study and monitoring of marine ecosystems. Its successful use in experiments underlines its potential for wider applications in marine science. As the system continues to evolve, further improvements through additional training and model refinement are likely to increase its effectiveness, making it an indispensable resource in the effort to protect and conserve the world's oceans.

A structural bioinformatics approach to study the effects of site-specific variations in Eosinophilic Esophagitis (EoE)

Deborah Giordano (1), Antonio d'Acierno (1), Anna Marabotti (2), Paola Iovino (3), Giuseppe Iacomino (1), Angelo Facchiano (1)

(1)National Research Council, Institute of Food Science, 83100 Avellino, Italy
(2) Department of Chemistry and Biology "A. Zambelli", University of Salerno, 84084 Fisciano, Italy
(3) Department of Medicine, Surgery and Dentistry "Scuola Medica Salernitana", University of Salerno, 84081 Baronissi, Italy.

In order to develop a bioinformatics resource to support the study of the molecular basis of eosinophilic esophagitis, a rare chronic immune/antigen-mediated inflammatory disorder, we studied the structure of the eotaxin-3 protein and the potential effects of known variations reported in public databases. Eotaxin-3 is a potent activator of eosinophil migration and activation, potentially leading to allergic airway inflammation. We applied a procedure established in our laboratory for previous studies (1, 2) to eotaxin-3 , firstly creating a complete 3D model of the protein and then simulating the structure of 105 protein variants resulting from known point mutations. The effects of amino acid substitutions on the protein's structure, stability, and potential function were detected and described in detail using bioinformatic analysis. Moreover, to offer our results in form of a resource to scientists interested in this field, we developed a web application to browse the analysis results and visualize the 3D models, allowing users to download the models for further analysis with their own software (3). Among the 105 amino acid substitutions investigated, the study identified at least one change in any of the examined structural parameters in 44 cases. Six other variations, while not showing detectable structural effects in our analysis, involve highly conserved amino acids, suggesting a possible functional role. All these variations require particular attention, as they may lead to a loss of protein functionality.

REFERENCES

1) d'Acierno, A., Scafuri, B., Facchiano, A., Marabotti, A. The evolution of a Web resource: The Galactosemia Proteins Database 2.0. Hum. Mutat. 2018, 39, 52–60.

2) Biancaniello, C., D'Argenio, A., Giordano, D., Dotolo, S., Scafuri, B., Marabotti, A., d'Acierno, A., Tagliaferri, R., Facchiano, A. Investigating the Effects of Amino Acid Variations in Human Menin. Molecules 2022, 27, 1747.

3) Giordano, D., d'Acierno, A., Marabotti, A., Iovino, P., Iacomino, G., Facchiano, A. Bioinformatics Study on Site-Specific Variations of Eotaxin-3, a Key Chemokine in Eosinophilic Esophagitis (EoE). Genes (Basel). 2024 15, 1073.

Automated cellular imaging to investigate the role of microtubule regulators during neuronal cell differentiation over time

Ludovica Altieri (1,2), Vincenzo Costanzo (1), Silvia Gasparini (1,3), Cecilia Mannironi (1), and Patrizia Lavia (1)

(1) Institute of Molecular Biology and Pathology (IBPM), CNR National Research Council, Rome, Italy
(2) Department of Biology and Biotechnology "Charles Darwin", Sapienza University of Rome, Italy
(3) Fondazione Telethon, Rome, Italy

Cellular imaging is rapidly evolving from a descriptive to a quantitative tool, aiming to overcome two major challenges: i) resolving the dynamic dimension of biological processes over time, and ii) capturing the heterogeneity among cells, which, even when genetically identical, can modulate their response to differentiating, developmental or mitogenic stimuli. Here we have used advanced microscopy videorecording methods to gain information on the process of neurodifferentiation at the single cell level over time. Microtubule-based structures are crucial to the organisation of the central nervous system, and many mutations implicated in neurodevelopmental disorders (NDDs) fall in genes that share the common trait of encoding microtubule regulators. Among them, CENP-F is a large protein with well-characterized functions in regulation of mitotic microtubules. CENP-F is mutated in individuals affected by primary hereditary microcephaly, a severe NDD associated with reduced brain size, suggesting a critical role of this protein during brain development. Its specific role however remains elusive. To assess how CENP-F affects neurodifferentiation, we have used two neurodifferentiatable cellular models, i.e. metastatic neuroblastoma SKNBE(2) cells, that can re-differentiate towards a neuronal phenotype in the presence of specific inducers, and murine primary cortical neurons, that can differentiate in culture. We down-regulated CENP-F with small interfering RNAs (siRNAs) in both models, then videorecorded CENP-F-proficient or deficient cultures. Processing of the resulting complex files requires unbiased automated methods for image acquisition and analysis. Here we have optimised image acquisition protocols, image segmentation and subsequent classification by training an artificial intelligence-based algorithm, obtaining the following results: 1) we have pinpointed the critical time window during which differentiation-specific processes, i.e. the appearance of neuritic extensions and the organisation of cellular connections, become quantifiable; 2) we have mapped their kinetics over several days; 3) we have identified steps of differentiation impaired by CENP-F silencing, including: maturation of cortical neurons, organisation of the neuronal network, neurite polarisation, and astrocyte expansion. Thus, automated videorecording assisted by machine-learning provide highly informative tools to dissect the role(s) of candidate genes in cellular models, helping to expand our understanding of the bases of complex developmental disorders.

Polygenic scoring methods in complex traits

Luigi Casillo (1), Francesca Bonometti (2), Mirko Treccani (3), Dramane Dagnogo (4), Giovanni Malerba (3)

(1) Department of Diagnostics and Public Health, University of Verona, 37134 Verona, Italy
(2) Innovagenome S.r.l., 25125 Brescia, Italy
(3) Department of Surgery, Dentistry, Paediatrics and Gynaecology, University of Verona, 37134 Verona, Italy
(4) Department of Integrative Cellular and Computational Biology, University of Trento, 38123 Trento, Italy

Complex traits depend on several known and unknown components, that are both genetic and environmental, and their interactions. Many genes contribute to the genetic component, each having a different, usually small, effect on trait variability. This makes it difficult to identify and combine genetic components in order to predict trait liability. The Polygenic Risk Score (PRS) is a predictive model that provides a score for a trait, according to the individual's genotype profile. The classic PRS consists of a weighted sum of risk alleles from different loci. It is developed using the Clumping + Thresholding (C+T) method and it takes advantage of risk alleles identified by Genome Wide Association Studies (GWASs). New PRS methods are being explored, including different Machine Learning algorithms, and among them, Random Forests (RFs). Random Forests (RFs) is a supervised machine learning algorithm suited for problems where the number of observations is way smaller than the number of variables, as usually the case of genetic studies. A trained RFs model can be used as a PRS since it provides a score for a trait, according to the individual's genotype profile. At present, a comprehensive comparison between the RFs and the classic PRS is missing. In order to evaluate and compare the performance of RFs and classic PRS, the two models were developed and tested on real data-based simulated genotyped individuals, under different complex realistic disease models. A population of 8,000 individuals genotyped at ~ 500,000 loci was simulated based on 1000 Genome Project data with disease models based on polygenic scores deposited on PGS Catalog. The simulated population was split into two sets, a 'GWAS set' used to perform a GWAS and a 'score set' used to tune and test the two methods. The tuning chose the best parameter values through a 10-fold cross validation procedure. The parameters included the GWAS p-values, the number of trees and mtry (the number of variables randomly sampled at each node of the trees) for the classic PRS and the RFs, respectively. Performance was evaluated through the value of the Area Under the Receiver-Operator Curve (AUC-ROC) and inspection of scores distributions in simulated cases and controls. Moreover, the computational demand was considered and reported. RFs displayed higher AUC-ROC values in all scenarios, compared to the classic PRS. Additionally, the distributions of scores in cases and controls were different and the difference was sharper for RFs than classic PRS. The RFs tuning was computationally more demanding than classic PRS, both in terms of RAM usage and time. In conclusion, our preliminary results suggest that RFs could be a more powerful tool for polygenic scoring with respect to the classic and widely used classic PRS.

Drug repurposing approach by structure-based virtual screening to identify novel putative VCP/p97 inhibitors

Carmen Maccanico (1), Palmina Bagnara (1), Elena Di Gennaro (1), Alfredo Budillon (2), Susan Costantini (1)

(1) Experimental Pharmacology Unit, Laboratori Mercogliano, Istituto Nazionale Tumori - IRCCS - Fondazione G. Pascale, 80131 Napoli, Italy
(2) Scientific Directorate, Istituto Nazionale Tumori - IRCCS - Fondazione G. Pascale, 80131 Napoli, Italy

Drug repurposing, defined as researching new indications for already approved drugs, is a strategy that, by focusing on specific molecular targets, offers significant advantages in decreasing the development cost and time to market over standard discovery. In the last years, our group focused the research on drug repurposing with the aim to select new therapeutic strategies to improve the outcome of cancer patients mainly in those cancers when no effective treatments currently exist. Our strategy is based on the identification of the drug candidates on a specific target through computational and experimental approaches using different drug libraries (ZINC20, ChEMBL, DrugBank and others). Valosin-containing protein (VCP/p97), an ATPase participating in diverse cellular processes, is one target for drug repurposing identified by our group. This protein interacts with multiple cofactors, and is necessary for the endoplasmic-reticulum-associated protein degradation (ERAD) pathway to maintain protein homeostasis. Its high expression has been found in many cancers among which pancreatic and breast cancers. Currently, many efforts have been devoted to design novel potential molecules able to inhibit VCP/p97 activity. CB-5083 and CB-5339 were identified as potent VCP/p97 inhibitors with moderate oral bioavailability, being able to induce ER stress. Unfortunately, two phase I clinical trials evaluating CB-5083 in cancer patients were terminated due to adverse effects (NCT02243917 and NCT02223598). About the other inhibitor, CB-5339, one clinical trial was withdrawn (NCT04372641) and one was completed even if no results were still posted (NCT04402541). In this context, drug repurposing strategy can represent an attractive approach to accelerate the selection of novel safe and effective VCP/p97 inhibitors at a lower cost and in a shorter time. We apply a structure-based virtual screening approach to identify FDA-approved drugs and commonly used for therapeutic purpose, that might be able to bind VCP/p97 and prevent the ATP binding, thus blocking its activity. In detail, we conducted a virtual screening of VCP using as target its three-dimensional structure and as ligands the world-approved drugs in ZINC20 library, by Autodock4.2 and Autodock Vina 1.2.5 docking tools. Both methods enabled the identification of some candidate drugs with favorable binding affinities to the VCP active site. We are currently testing the ability of the selected drugs to inhibit VCP/p97 ATPase activity as well as cell proliferation on a panel of breast and pancreatic cancer cell lines in order to confirm their antitumor potential.

FACE2FACE: a simple tool for macromolecule interface analysis

Mario Incarnato (1), Gianmarco Pascarella (1), Allegra Via (2), Veronica Morea (1), Patrizio Di Micco(1)

(1) CNR Institute of Molecular Biology and Pathology, Rome, IT;

(2) Sapienza University of Rome, Dept. of Biochemical Sciences "A. Rossi Fanelli", Rome, IT

Motivation: Analysis of interfaces in macromolecular complexes is essential to unveil the mechanisms underlying molecular recognition. However, currently available interface analysis tools are focused on specific macromolecules (generally proteins) and output information and format.
Results: We have developed a novel fast and comprehensive tool for macromolecule interface analysis. The program provides information about proteins or nucleic acids interactions with other biological macromolecules and/or small molecules, in formats that can be easily parsed and ready to be imported in spreadsheet applications and in widely used structure visualization and analysis programs such as PyMol and ChimeraX.
Availability: FACE2FACE is available at http:/face2face.ibpm.cnr.it/.

The Role of Amino Acid Metabolism in Autosomal Dominant Polycystic Kidney Disease Progression

Gözde Ertürk Zararsız [1,2], Serra İlayda Yerlitaş [1,2], Ahu Cephe [3], Alparslan Demiray [4], Salih Güntuğ Özaytürk [4], Necla Kochan [5], Halef Okan Doğan [6,7], Gökmen Zararsız [1,2,7], İsmail Koçyiğit [4]

[1] Department of Biostatistics, School of Medicine, Erciyes University, Kayseri, Türkiye
[2] Drug Application and Development Research Center, Erciyes University, Kayseri, Türkiye
[3] Institutional Data Management and Analytics Coordination Unit, Erciyes University, Kayseri, Türkiye
[4] Department of Nephrology, School of Medicine, Erciyes University, Kayseri, Türkiye
[5] Department of Mathematics, Izmir Economy University, İzmir, Türkiye
[6] Department of Biochemistry, School of Medicine, Altinbas University, İstanbul, Türkiye
[7] Erciyes Teknopark, Hematainer Biotechnology and Health Products Inc. Kayseri, Türkiye

Introduction: Amino acid metabolism is crucial for human health and well-being, as amino acids serve as the fundamental components of proteins and fulfill various essential functions in the body [1-2]. The aim of this study was to perform quantification of serum amino acid levels in Autosomal Dominant Polycystic Kidney Disease (ADPKD) patients and to compare them according to clinical outcomes of disease progression (slow and rapid progression), development of hypertension (present/absent) and mortality (ex and survived).
Methods: Serum blood samples for metabolomic analyses were obtained from 254 patients who were diagnosed with ADPKD in Erciyes University Faculty of Medicine, Nephrology Clinic, who were still being followed up during the study period. The serum amino acid profile was determined using liquid chromatography Tandem Mass Spectrometry (LC-MS/MS).
Results: Isoleucyl-Proline and L-Glutamine metabolic profiles in patients with rapid progression of ADPKD differ significantly from those in the slow progression group. Isoleucyl-Proline levels were elevated in the rapid progression group, whereas L-Glutamine levels were higher in the slow progression group. Significant differences were observed in the levels of 4-Methylaminobutyrate, Acetyl-L-arginine, Alanine, Isoleucyl-proline, L-Carnitine, and L-Tryptophan in patients who died. Additionally, significant differences were found in the levels of Acetyl-L-arginine, Asparagine, Hydroxyarginine, Isoleucine, Isoleucyl-proline, and L-Lysine in patients with hypertension compared to those without hypertension. Distinct metabolic reprogramming is evident, particularly involving alterations in alanine and asparagine metabolism. These metabolic shifts are closely linked to the molecular mechanisms driving the disease. Furthermore, statistically significant associations were identified between specific amino acids levels.
Conclusion: Investigating innovative therapeutic and diagnostic approaches targeting specific amino acids could provide promising avenues for reducing disease severity. These findings enhance our comprehension of the complex interplay between amino acid metabolism and overall health, laying the groundwork for future studies focused on advancing disease detection and management strategies.

Literature:
[1] Wu G. Functional amino acids in nutrition and health. Amino Acids. 2013;45(3):407-11. doi: 10.1007/s00726-013-1500-6.
[2] Ishii I, & Bhatia M. Amino Acids in Health and Disease: The Good, the Bad, and the Ugly. Int J Mol Sci. 2023;24(5):4931. doi: 10.3390/ijms24054931.

(This study was supported by the Research Fund of Erciyes University [TSAÜ-2024-13479].)

Fatty Acid Metabolism in the Progression of Autosomal Dominant Polycystic Kidney Disease

Gökmen Zararsız [1,2,3], Gözde Ertürk Zararsız [1,2], Serra İlayda Yerlitaş [1,2], Ahu Cephe [4], Alparslan Demiray [5], Salih Güntuğ Özaytürk [5], Necla Kochan [6], Halef Okan Doğan [3,7], İsmail Koçyiğit [5]

[1] Department of Biostatistics, School of Medicine, Erciyes University, Kayseri, Türkiye
[2] Drug Application and Development Research Center, Erciyes University, Kayseri, Türkiye
[3] Erciyes Teknopark, Hematainer Biotechnology and Health Products Inc. Kayseri, Türkiye
[4] Institutional Data Management and Analytics Coordination Unit, Erciyes University, Kayseri, Türkiye
[5] Department of Nephrology, School of Medicine, Erciyes University, Kayseri, Türkiye
[6] Department of Mathematics, Izmir Economy University, İzmir, Türkiye
[7] Department of Biochemistry, School of Medicine, Altinbas University, İstanbul, Türkiye

Introduction: Recent developments in metabolomics have highlighted the intricate nature of fatty acid metabolism in health and disease. As a crucial class of metabolites, changes in free fatty acid (FFA) levels are significantly involved in the development and progression of various diseases. The aim of this study was to perform quantification of serum fatty acid levels in Autosomal Dominant Polycystic Kidney Disease (ADPKD) patients and to compare them according to clinical outcomes of disease progression (slow and rapid progression), development of hypertension (present/absent) and mortality (ex and survived).
Methods: Plasma samples for metabolomic analysis were collected from 254 patients diagnosed with ADPKD at the Nephrology Clinic of Erciyes University Faculty of Medicine. These patients were being actively monitored during the study period. The serum amino acid profiles were analyzed using Liquid Chromatography Tandem Mass Spectrometry (LC-MS/MS).
Results: No differences were observed in fatty acid metabolic profiles between patients with rapid progression of ADPKD and those with slower disease progression. However, significant differences in LysoPE(0:0/18:2(9Z,12Z)) and PE(14:1(9Z)/16:1(9Z)) profiles were observed between mortality groups. In addition, significant differences were found in the levels of LysoPC(0:0/18:0), LysoPE(0:0/18:0), LysoPE(0:0/18:3(6Z,9Z,12Z)), and LysoPE(0:0/20:0) between patients with hypertension and those without hypertension. These metabolic alterations are strongly associated with the underlying molecular mechanisms of ADPKD. Additionally, statistically significant correlations were found between specific fatty acid levels.
Conclusion: The findings of this study emphasize the therapeutic potential of fatty acid metabolism in ADPKD and underline the prognostic value of fatty acids in these diseases.

Literature:
[1] Kazantzis M, Stahl A. Fatty acid transport proteins, implications in physiology and disease. Biochim Biophys Acta. 2012;1821(5):852-7. doi: 10.1016/j.bbalip.2011.09.010.

(This study was supported by the Research Fund of Erciyes University [TSAÜ-2024-13479].)

Machine-Learning-Based Modeling of Amino Acid Levels in Crimean-Congo Hemorrhagic Fever Patients

Ahu Cephe [1], Seyit Ali Büyüktuna [2], Serra İlayda Yerlitaş [3,4], Gözde Ertük Zararsız [3,4], Kübra Doğan [5], Demet Kablan [6], Gökhan Bağcı [7], Selda Özer [6], Cihad Baysal [2], Yasemin Çakır [2], Necla Koçhan [8], Halef Okan Doğan [6,9], Gökmen Zararsız [3,4,9]

[1] Institutional Data Management and Analytics Coordination Unit, Erciyes University, Kayseri, Türkiye
[2] Department of Infectious Disease and Clinical Microbiology, Cumhuriyet University School of Medicine, Sivas, Türkiye
[3] Department of Biostatistics, School of Medicine, Erciyes University, Kayseri, Türkiye
[4] Drug Application and Development Research Center, Erciyes University, Kayseri, Türkiye
[5] Department of Biochemistry, Minister of Health Sivas Numan Hospital, Sivas, Türkiye
[6] Department of Biochemistry, Cumhuriyet University School of Medicine, Sivas, Türkiye
[7] Department of Biochemistry, School of Medicine, Altinbas University, İstanbul, Türkiye
[8] Department of Mathematics, Izmir University of Economics, İzmir, Türkiye
[9] Erciyes Teknopark, Hematainer Biotechnology and Health Products Inc. Kayseri, Türkiye

Aims: Crimean-Congo Hemorrhagic Fever (CCHF) is one of the most prevalent tick-borne infections and is associated with high mortality rates. This disease, known for its potential to cause epidemics, poses significant diagnostic challenges due to its variable progression characteristics. Additionally, there is currently no antiviral treatment available. Thus, identifying biomarkers for disease progression and predicting outcomes early is crucial for patient survival. This study aims to predict the amino acid profile in patients with Crimean-Congo Hemorrhagic Fever using machine learning models.

Materials and Methods: We analyzed amino acid data from 190 individuals, comprising 115 CCHF patients, 30 CCHF-negative patients, and 45 healthy controls. A total of 32 amino acid variables were utilized as predictors. To reduce the number of predictor variables, we employed variable selection methods, including Boruta and recursive feature elimination. We applied several algorithms—Random Forest (RF), Support Vector Machine (SVM), k-Nearest Neighbors (KNN), Least Absolute Shrinkage and Selection Operator (LASSO), Nearest Shrunken Centroids (NSC), and eXtreme Gradient Boosting (XGBoost)—to predict the amino acid profiles. Analyses were conducted using 10-fold cross-validation with 5 repeats across 30 different training and test sets, all implemented in R. Model performance was evaluated using metrics such as AUC (Area Under Curve), sensitivity, specificity, accuracy, positive predictive value, negative predictive value, positive likelihood ratio, and negative likelihood ratio. The varImp function from the caret package was used to identify the most important predictor variables for the amino acid profiling model.

Results: Our results indicated that machine learning algorithms yielded superior outcomes, particularly the models achieving the highest AUC values. The Lasso model with Boruta feature selection, achieved the highest AUC value (AUC=0.96). This was followed by XGBoost algorithms using Boruta feature selection (AUC=0.95), Random Forest using Boruta feature selection (AUC=0.94), Lasso algorithms (AUC=0.94) employing recursive feature elimination, and NSC with Boruta feature selection (AUC=0.93).

Conclusion: These findings suggest that machine learning can be an invaluable tool for understanding amino acid profiles in Crimean-Congo Hemorrhagic Fever.