BIOINFORMATICS
OPEN X DAYS

>CONFERENCE_BOOK

P-BIO elixir NEBIUM BioData.pt

WELCOMING MESSAGE

Greetings,

On behalf of the Organizing Committee, we are pleased to welcome you to the 10th edition of Bioinformatics Open Days. It is an honor to have a diversity of students, researchers, and academics, attending our event and supporting this endeavor to shine a light on some of the exceptional work that has been produced in the field of bioinformatics.

Bioinformatics has been significantly growing both as a technological and a scientific field, providing, in its own right, enormous challenges, and opportunities both in research and for companies. Bioinformatics in Portugal, similarly to the worldwide scenario, has experienced outstanding growth over the past few years. Such is reflected academically, through the development of prestigious post-graduations, and economically, or in the business sector, with the establishment of new start-up companies with international connections.

Bioinformatics Open Days is a student-led initiative, first held in 2012, at the University of Minho, Braga. It aims to promote the exchange of knowledge between students, teachers, and researchers from the Bioinformatics and Computational Biology fields. Each year, this event aims to describe the present and the future of Bioinformatics, both nationally as internationally. For this year's 10th edition, we are pleased to host an entire day dedicated to "Unleash our potential to shape the future!" organized by ELIXIR Portugal to celebrate their 5th anniversary.

This year, the atypical circumstances had forced us to hold our event in an online format. We maintained our commitment to provide you an event with such quality as the Bioinformatics Open Days have offered you over the past 10 years.

Finally, we are pleased to welcome all the participants to this event. We hope everyone has great attendance where, hopefully, you can share valuable knowledge.

The Organizing Committee of BOD 2021

THE ORGANIZING COMMITTEE

Miguel Rocha
General Chair

Fernanda Vieira

Maria João Lopes

Laura Duro

Vasco Silva

Carina Afonso

Miguel O. Rocha

Ana Barbosa

Joana Ribeiro

João Monteiro

Maria Faria

Miguel Martins

Cláudio Monteiro

Gil Afonso

Maria Couto

Paulo Carvalhais

Cátia Mendes

Tiago Machado

Francisco Pereira

Sofia de Beir

Miguel Pacheco

José Duarte

Mariana Coelho

PROGRAM

May 5th - Wednesday

11:30 h	Opening session Mário Gaspar da Silva
11:45 h	Panel I: *Bioinformatics and Biological Data Management* Chaired by **Mário Gaspar da Silva** *Insights from the State-of-the-art and a Look to the Future* - **Niklas Blomberg** *Tracked Roads and Roads to Track* - **Arlindo Oliveira**
13:00 h	Lunch
14:30 h	Panel II: *Training in Bioinformatics and Biological Data Management* Chaired by **Cátia Pesquita** *Teaching Bioinformatics and Biological Data Management in Portugal - a vision from 2026* - **Miguel Rocha** *Training in Bioinformatics and Biological Data Management worldwide - a vision from 2026* - **Celia van Gelder**
15:45 h	Round table: *The ELIXIR-PT Data Framework and Domain-specific Data Challenges* Chaired by **Ana Teresa Freitas** Health Data - **Carla Oliveira** Marine Resources - **Cymon Cox** Industrial Microbiology - **Isabel Rocha** New Drugs - **Irina Moreira** Plant Sciences - **Nelson Saibo**
17:00 h	Closing Remarks José Pereira Leal

May 6th - Thursday

09:30 h	Opening session
10:00 h	Keynote Lecture *Towards the integration of common and rare variant analysis in genotype-to-phenotype studies* - Pedro Beltrão
10:45 h	Break
11:00 h	Oral Communications - Session 1 Chaired by Miguel Rocha, University of Minho ***Structural Bioinformatics***
11:00 h	*Structure-based virtual screening, Molecular Dynamics and free energy calculations for the identification of novel inhibitors against biofilm formation by C.violaceum* - Fábio Martins
11:15 h	*Protein Engineering of Mini Therapeutic Proteins against SARS-CoV-2* - Carlos Cruz
11:30 h	*Gibbs Free Energy Variation Analysis of Breakpoint Regions from MtDNA Human Brain Deletions* - João Carneiro
11:45 h	*MD simulations reveal that the parainfluenza FP forms oligomeric pore-like structures inside a membrane* - Mariana Valério
12:00 h	*New insights on SARS-Cov-2 M protein: Structure, Member Orientation and Mutations* - Catarina Marques-Pereira
12:15 h	*Deciphering the interaction of lactoferrin with V-ATPase towards a deeper understanding of its mechanisms of action* - Cátia Santos-Pereira
13:00 h	Lunch
14:30 h	Keynote Lecture *Integrating metabolomics, cheminformatics, and metabolic modeling for the discovery of new metabolic pathways in KBase* - Christopher S. Henry

(Continues on the next page)

15:15 h	Oral Communications - Session 2 Chaired by **Vitor Pereira, University of Minho** ***Systems Biology, Omics Data and Machine Learning***
15:15 h	*Revising lipid chemical structures in genome-wide metabolic models with BOIMMG* - **João Capela**
15:30 h	*Dynamic genome-scale modelling of the Saccharomyces non-cerevisiae yeasts metabolism in wine fermentation* - **David Santos**
15:45 h	*Towards an automatic cross-species comparative genomics portal* - **Jorge Oliveira**
16:00 h	*A comparative evaluation of dimensionality reduction methods on large-scale gene expression datasets* - **Sara Ribeiro**
16:15 h	*BioTMPy: a Deep Learning-based tool to classify biomedical literature* - **Nuno Alves**
16:45 h	Break
17:00 h	Poster Communications - Session 1 *Biotechnology applications, Structural Bioinformatics and Systems Biology*
18:00 h	Break
18:15 h	Quiz

PROGRAM

May 7th - Friday

09:30 h	**Keynote Lecture** *Computational Analyses of cancer genomes* - **Núria Lopez-Bigas**
10:15 h	**Poster Communications - Session 2** *Omics Data Analysis, Machine Learning and Health Applications*
11:15 h	**Break**
11:30 h	**Oral Communications - Session 3** Chaired by **Oscar Dias, University of Minho** ***Genomics and Health***
11:30 h	*Single-cell genomics in oligodendroglia: opening doors to understand multiple sclerosis* - **Ana Mendanha Falcão**
11:45 h	*INSaFLU: an open web-based bioinformatics suite for influenza and SARS-CoV-2 genome-based surveillance* - **Miguel Pinheiro**
12:00 h	*SynPred: Prediction of Drug Combination 1 Effects in Cancer using Full-Agreement Synergy Metrics and Deep Learning* - **António Preto**
12:15 h	*One bioinformatics pipeline for genomic characterisation of Candida sp. clinical strains* - **Maria João Carvalho**
12:30 h	*Bioinformatic analysis reveals protein-based biomarkers for non-invasive Prostate Cancer detection* - **Tânia Lima**
12:45 h	*Complete genome sequences of sixteen Xanthomonas spp. strains assembled with short and long reads* - **Miguel Teixeira**
13:00 h	**Lunch**
14:30 h	**Network Session** CBR Genomics HeartGenetics iLoF P-Bio SilicoLife
16:00 h	**Break**
16:15 h	**Round table: *Studying Bioinformatics*** Chaired by **Miguel Martins** **Ana Carolina Rodrigues** - *University of Lisbon* **João Capela** - *University of Minho* **Manuel Pires** - *University of Porto*
17:00 h	**Closing Session**

May 8th - Saturday

09:30 h	Workshops *An introduction to Chemoinformatics Workshop* - João Capela, João Correia, Tiago Sousa *Artificial Intelligence in Drug Discovery Workshop* - Irina Moreira *Computational Protein Design Workshop* - Carlos Cruz, Diana Lousa

ELIXIR Portugal is completing 5 years of excellence in supporting research, training, and innovation in the national scientific system through the management and advanced analysis of biological data. To celebrate this occurring, ELIXIR Portugal has decided to organize a dedicated day to showcase what has been done in the past few years but also to foster an active discussion regarding the increasingly emerging fields of bioinformatics and data management.

Panel I: *Bioinformatics and Biological Data Management*

Mário Gaspar da Silva

Mário Gaspar da Silva is ELIXIR Portugal's Head of Node. He is also a Full Professor at the Instituto Superior Técnico (IST) and a Senior Researcher at the Instituto de Engenharia de Sistemas e Computadores - Investigação e Desenvolvimento (INESC-ID). Mário capitalizes on more than 30 years of research in web engineering and natural language processing, biomedical informatics, and digital libraries.

Niklas Blomberg

Niklas Blomberg is ELIXIR's Executive Director. Before joining ELIXIR in 2013, he worked as a Principal Scientist and Team Leader in Computational Chemistry and Computational Biology at AstraZeneca (1999-2013). Niklas has a lot of experience in coordinating EU H2020 projects. In the past, he has coordinated ELIXIR-EXCELERATE, a project to develop the ELIXIR infrastructure, and CORBEL. Currently, he is coordinating EOSC-Life, which aims to build a collaborative digital space for European life science research, and ELIXIR-CONVERGE, a project to help standardize life science data management across Europe.

Arlindo Oliveira

Arlindo Oliveira is a Distinguished Professor at Instituto Superior Técnico (IST). Arlindo is the former president of IST and current president of INESC. His research focuses mainly on algorithms and complexity, machine learning, bioinformatics, and digital circuit design and has contributed hugely to the progress of the collaboration between life sciences and computation, crucial in the time being.

Panel II: *Training in Bioinformatics and Biological Data Management*

Cátia Pesquita

Cátia Pesquita is an Assistant Professor at the Faculdade de Ciências da Universidade de Lisboa (FCUL). Cátia is a researcher at LASIGE where she leads the research line of excellence in health and biomedical informatics. Her main interest is to "turn data into meaningful and purposeful knowledge" and for that she focuses her research mostly on biomedical ontologies, semantic web, ontology matching, semantic similarity, ontology evolution, knowledge management, and data mining

Miguel Rocha

Miguel Rocha is an Associate Professor at the Universidade do Minho. Miguel leads the Bioinformatics and Systems Biology research group at the Centro de Engenharia Biológica (CEB) and teaches subjects related to bioinformatics, natural computation, data mining/machine learning at the Informatics Department of Universidade do Minho being also the director of the MSc in Bioinformatics, a degree he co-founded in 2007.

Celia van Gelder

Celia van Gelder is a Training Programme Manager at the Dutch Techcentre for Life Sciences (DTL). Celia is a manager, coordinator, and trainer with extensive experience in initiating and managing bioinformatics education & training projects at institute, national and international levels. She also co-leads the ELIXIR Training Platform, which intends to build a sustainable training infrastructure in Europe.

Round table: *The ELIXIR-PT Data Framework and Domain-specific Data Challenges*

Ana Teresa Freitas

Ana Teresa Freitas is HeartGenetics's Executive Director. Ana Teresa is also a Full Professor at the Instituto Superior Técnico (IST) and a Senior Researcher at the Instituto de Engenharia de Sistemas e Computadores - Investigação e Desenvolvimento (INESC-ID) in the Knowledge Discovery and Bioinformatics group focusing her research in computational biology, algorithms and complexity, and data mining.

Carla Oliveira

Carla Oliveira is the leader of the Expression Regulation in Cancer group at IPATIMUP and an Associate Professor at the Faculdade de Medicina da Universidade do Porto. Carla's research focuses on the identification of cancer biomarker signatures for diagnostics, treatment, and treatment resistance, using next-generation sequencing (NGS), bioinformatics, cell line, and animal models. Her team has a strong link with hospitals that provide patients' biological material, and the industry that co-funds the research. Besides, she launched the Bioinf2Bio bioinformatics company, in August 2013.

Cymon Cox

Cymon Cox is the leader of the Plant Systematics and Bioinformatics group at CCMAR and the coordinator of the Marine Resources Community at BioData.pt. As a phylogenetic systematist, Cymon's main research aim is the reconstruction of ancestral relationships among organisms using molecular and morphological character data.

Isabel Rocha

Isabel Rocha is the leader of the Systems and Synthetic Biology group at ITQB-NOVA and the pro-rector of the NOVA University of Lisbon. She is also the scientific coordinator, among others, of the Shikifactory100 project and the national representative on the executive board of ELIXIR. Her research in biotechnology covers the topics of bioinformatics, metabolic engineering, synthetic biology, and systems biology, and she has published more than 140 articles in international magazines, books, and conferences.

Irina Moreira

Irina Moreira is an Associate Professor at the Centre for Neuroscience and Cell Biology (CNC-UC) where she leads the Data-Driven Molecular Design research group. She is interested in exploring the interface between computer sciences and structural biology through "chemistry, bioinformatics, biophysics, and data mining to develop and implement novel approaches and address emerging biological questions that cannot be tackled with traditional approaches alone".

Nelson Saibo

Nelson Saibo is the leader of the Plant Gene Regulation Lab at ITQB NOVA where he uses model and crop plants to study gene regulatory mechanisms underlying plant growth and plant responses to adverse environmental conditions. He has over 15 years experience in plant molecular biology which makes him the perfect researcher to bring us the diverse (and huge!) challenges raised by plant-associated data.

Closing Remarks

José Pereira Leal

José Leal is a bioinformatician with close to 20 years of professional experience. After 10 years abroad, he returned to Portugal where he established and led the Computational Genomics Laboratory at the Instituto Gulbenkian de Ciência (IGC), a group focused on comparative genomics, medical genomics, and data integration. He also coordinated the #Bioinformatics Unit of IGC, where he helped to establish ELIXIR, and its national counterpart, BioData.pt. His commitment to translating the latest advances in Genomic and Bioinformatics research to improve clinical practice and patient care has contributed hugely to the progress of the field in Portugal.

Christopher S. Henry

Christopher Henry is a computational biologist at the mathematics and computer science division of Argonne National Laboratory (ANL). He obtained his PhD studying Biochemical Thermodynamics, Genome-Scale Metabolic Modelling and automatically generating novel biochemistry at Northwestern University. His primary research focuses on the prediction of phenotype from genome through the use of comparative genomics, metabolic modeling, and dynamic cellular community models, and his team is involved in several projects with the University of Chicago.

Núria Lopez-Bigas

Núria Lopez-Bigas is an ICREA Researcher Professor at the Institute for Research and Biomedicine (IRB) in Barcelona where she also leads the Biomedical Genomics Research Group. Besides her PhD in Biology, Núria has expertise in Medical Genetics, Computational Biology, and Bioinformatics. Her research concentrates on the study of cancer from a genomics perspective. Particularly, her research group attempts to understand mutational processes finding cancer drivers (such as mutations, genes, and pathways) and contributing to precision cancer medicine. Her lab had important achievements such as the development of IntOGen, which is a discovery tool for cancer research, and pioneer methods to identify driver genes (Oncodrive methods).

Pedro Beltrão

Pedro Beltrão is the leader of *the Evolution of Cellular Networks* group at EMBL-EBI. In the scope of his Biology PhD at the University of Aveiro, he researched at EMBL-Heidelberg and further conducted postdoctoral research at the University of California, San Francisco. His studies concentrated on the divergence of cellular functions during evolution and how these divergences steer differently in pathogenic pathways, and assistance in data analysis on the SARS-CoV-2 virus. Pedro's research group is developing a general propose framework to predict the molecular consequences of DNA changes and using these to guide genotype-phenotype associations.

Ana Carolina Rodrigues

Ana Carolina Rodrigues is a master's student of Bioinformatics and Computational Biology at the Faculty of Sciences of the University of Lisbon (FCUL). During her bachelor's degree in Cellular and Molecular Biology at FCT NOVA, she interned at the Microbial Genetics lab, under Professor Isabel Sá Nogueira's orientation. Ana Carolina is completing her master's dissertation in the Papanikolaou Lab (Computational Clinical Imaging Group) at the Champalimaud Foundation where she works with machine learning to develop a classifier that predicts prostate tumor aggressiveness from MRI images.

João Capela

João Capela holds a bachelor's degree in biology and a master's degree in Bioinformatics from University of Minho. He has developed a repository for lipid chemical structures and tools for establishing relevant biosynthetic relationships between those structures and further integration in Genome-scale Metabolic Models (GSMM). Moreover, since 2019, João has been involved in the continuous development of merlin, a large-scale software to assist the reconstruction of GSMM. Currently, João is a research fellow in the DeepBio project at the University of Minho, where he is developing a Machine Learning (ML)-oriented database and improving ML and Deep Learning models for the prediction of new sweeteners.

Manuel Pires

Manuel Pires is a master's student of Bioinformatics and Computational Biology at the University of Porto. Graduated in Biochemistry from the University of Coimbra, Manuel has been in contact with a wide range of areas including Bioinformatics, Data Science, and Machine Learning. He has been in contact with programming and chemoinformatic areas during his bachelor's project at Moreira Lab where he is currently working in the scope of his master's dissertation, under Professor Irina Moreira's orientation.

CBR Genomics

CBR Genomics is a Portuguese technological Start-Up that merges IT and Biotech, whose mission is to promote the usage of genomic data in the clinical practice. It develops genetic services that screen for hundreds of genetic diseases and, through its patented technology, provide clinical reports containing relevant genomic information for patients' health management. At CBR Genomics we work daily to foster the usage of genetic information in the clinical practice, so that it will become as ordinary as x-ray or blood testing, thus enabling the change of Medicine's Paradigm towards a more Predictive and Preventive approach.

HeartGenetics

HeartGenetics is a digital health company that combines intelligence with genetic data. It aims to apply genetic knowledge and testing as an instrument in the definition of highly personalized lifestyle plans. The company accommodates a multidisciplinary team of Genetics, Bioinformatics, and Molecular Biology specialists responsible for the development of wellness genetic tests. These tests supported by a strong knowledge of cardiovascular genetics are used to personalize lifestyle plans.

iLoF

iLoF is empowering a new era of personalized medicine using AI and photonics to build a cloud-based library of diseases biomarkers and biological profiles, likewise, providing novel technologies for screening and stratification in a quick, portable and affordable way. iLoF's platform mainly focuses on Alzheimer's treatment, but the company intends to expand the power of its technologies to other diseases such as Digestive Cancer, Stroke, and Infectious diseases.

P-BIO

P-BIO contributes to the development and support of biotechnology in Portugal. It is a unique association that comprehends the majority of companies from the Biotechnology and Lifescience sectors. P-BIO aims to establish a favorable environment for the creation and growth of start-ups, as well as their commercial expansion both nationally as internationally. Member of the EuropaBio, P-BIO connects their companies with different key partners, as the government, researchers, regulatory agencies, and other industry-connected institutions.

SilicoLife

SilicoLife designs optimized microorganisms and novel pathways for industrial biotechnology applications. The team includes specialists in several areas, namely biotechnology, computational biology, metabolic engineering, molecular biology, systems biology, bioinformatics, and text mining. SilicoLife builds computational models of microbial cells and develops proprietary state-of-the-art algorithms to find the most efficient pathways between raw materials and end-products, streamlining the strain design process and explore non-intuitive pathway modifications.

An introduction to Chemoinformatics Workshop

João Capela | João Correia | Tiago Sousa

Chemoinformatics (often referred to as chemical informatics or cheminformatics) is a field of study that combines chemistry, computer, and information sciences. In addition to inferring simple information like systematic names and chemical formulas, primary applications of this field include the storage, indexation, and mining of information related to chemical structures such as molecular functional groups, docking sites, chemical, and physical properties.

In this workshop, introductory topics related to the field of chemoinformatics will be explored using python. Firstly, various molecule computational representations will be approached, being followed by similarity and substructure matching techniques using molecular fingerprints, SMARTS notations, and Scaffolds. The workshop will also focus on analyzing large datasets of compounds resorting to both unsupervised and supervised exploration using Machine and Deep Learning techniques.

Artificial Intelligence in Drug Discovery Workshop

Irina Moreira

Data-driven Molecular Design: importance of data interpretation and pipeline interpretability on artificial intelligence targeting biological problems.

Computational Protein Design Workshop

Carlos H. Cruz | Diana Lousa

Designing proteins with optimal properties for a given objective is a long scientific quest, which has seen great advances in recent years.[1] As examples, protein scientists have been able to design fluorescent proteins with specific properties from scratch and develop potent protein inhibitors to fight COVID-19 and other diseases. For this, they used sophisticated algorithms that are able to design and/or engineer protein sequences to achieve the desired goal.

This workshop presents the basic concepts of structure-based computational protein design methods and provides a hands-on tutorial for designing proteins by motif-driven design techniques using the Rosetta software. The method consists in grafting a target motif (e.g. a domain that interacts with a partner protein) onto a scaffold protein (i.e. a protein with a stable structure). A scaffold library is used to select appropriate scaffold proteins, which are then used to graft the motif of interest. The next step is to redesign the interface and evaluate the best candidates by structural and thermodynamics metrics. The aim of the workshop is to introduce beginners to computational protein design methods using the Rosetta software.

1. Kuhlman B. and Bradley P., Advances in protein structure prediction and design, *Nature Rev. Mol. Cell Biol.*, **20**, 681-697 (2019)

Session 2
6th May
15:15 h

A comparative evaluation of dimensionality reduction methods on large-scale gene expression datasets

Ribeiro, S.[1,2], Ferreira, P.[1,2,3], Alves, C.[1], Ribeiro, R.[1]

[1] Faculty of Sciences of the University of Porto, Rua do Campo Alegre s/n, 4169-007 Porto.
[2] i3s/Ipatimup - R. Alfredo Allen 208, 4200-135 Porto.
[3] Inesc-Tec Liaad - Campus da FEUP, Rua Dr. Roberto Frias, 4200 - 465 Porto, Portugal.

Since the first draft of the human genome, the biggest challenge has been to interpret the structure and function of the genetic information. The typical high dimensionality of genomics data raises the necessity of efficient clustering and dimensionality reduction (DR) methods. They constitute crucial tools for the exploratory analysis of this data. It is essential to uncover structural or qualitative information of biological data in an intuitive way, by visualizing patterns and identifying irrelevant or highly correlated features. The transformation or removal of less informative data may also facilitate further data analysis and visualization by focusing on the dimensions or components that capture most of the variability and yet conserve the majority of the signal. There is currently a vast number of clustering and DR methods available. The selection of the most appropriate methods is dependent on dataset characteristics, such as the inherent data sparsity.

In this thesis, we propose a framework to apply, evaluate and compare, in a qualitative and quantitative way, different linear and non-linear DR methods in the analysis of gene expression datasets. The evaluation was performed on two publicly available and large-scale RNA-Seq gene expression datasets, including bulk and single-cell RNA-Seq data. A specific quantitative k-means clustering-based evaluation approach is proposed and complemented with several additional metrics.

Our results show that, at the cost of tuning several input hyperparameters, non-linear methods have an improved performance for both datasets, in particular local linear embedding (LLE), t-Distributed Stochastic Neighbour Embedding (t-SNE) and Uniform Manifold Approximation and Projection (UMAP). To alleviate this former difficulty, an hyperparameter-tuning approach is suggested. Finally, a summary of recommendations is provided, in order to help improve the usability and efficacy of these valuable methods.

Session 3
7th May
11:30 h

Bioinformatic analysis reveals protein-based biomarkers for non-invasive Prostate Cancer detection

Tânia Lima[1], Rui Henrique[2,3,4], Rui Vitorino[1,5], Margarida Fardilha[1]

[1] Laboratory of Signal Transduction, Department of Medical Sciences, Institute of Biomedicine iBiMED, University of Aveiro, 3810-193 Aveiro, Portugal
[2] Cancer Biology and Epigenetics Group, Research Center of Portuguese Oncology Institute of Porto (GEBC CI-IPOP) and Porto Comprehensive Cancer Center (P.CCC), 4200-072 Porto, Portugal
[3] Department of Pathology, Portuguese Oncology Institute of Porto (IPOP), 4200-072 Porto, Portugal
[4] Department of Pathology and Molecular Immunology, Institute of Biomedical Sciences Abel Salazar, University of Porto (ICBAS-UP), 4050-513 Porto, Portugal
[5] UnIC, Department of Surgery and Physiology, Faculty of Medicine, University of Porto, 4200-319 Porto, Portugal

Prostate cancer (PCa) is one of the most prevalent types of cancer. However, the limited accuracy and invasive nature of the currently used diagnostic tools (digital rectal examination, PSA serum levels, prostate biopsy) has driven the demand for new non-invasive biomarkers. Urine is a noninvasively collected biofluid that contains proteins that are secreted or have come in direct contact with the prostate, reflecting the molecular changes associate with this organ. Therefore, it is considered a valuable source of biomarkers. It is believed that the integration of proteomics data from different studies is vital for identifying new PCa biomarkers, but studies carried out in this regard have few converging results. Hence, using a different approach, the novelty of this study is the integration of urinary and tissue proteomes of PCa patients, focusing on urine-tissue overlaps. This comparative analysis increases the power of individual studies and places urinary proteome data as a reflection of the molecular changes occurring in PCa tissue. Considering that kidney and bladder are the main contributors to the urine proteome, the proteins expressed in these two organs, as well as in the tumors that affect them (kidney and bladder cancer), were not considered for the final list of potential urinary markers for PCa. A detailed bioinformatic analysis revealed molecular features consistently dysregulated in urine from PCa patients that mirror the alterations in prostate tumor tissue. Furthermore, MSMB, KLK3, ITIH4, ITIH2, HPX, GP2, APOA2 and AZU1 proteins stood out as candidate urinary biomarkers for PCa.

Session 2
6th May
15:15 h

BioTMPy: a Deep Learning-based tool to classify biomedical literature

Nuno Alves[1], Ruben Rodrigues[1], Miguel Rocha[1]

[1] BIOSYSTEMS, Centre of Biological Engineering, University of Minho, Campus de Gualtar 4710-057 Braga Portugal

Over the last few decades, the publication rate has been massively increasing, resulting in a huge number of available scientific documents, which consequently makes the search of relevant information for a certain topic a heavy and time-consuming task. Biomedical Text Mining has been addressing this problem for a while, but there is still space for improvements. For instance, PubMed, which contains now more than 32 million citations, has only recently implemented a machine learning model to improve document ranking, and is still trying to improve their system by implementing a Deep Learning model, needing for now further studies.

Following this line of thought, a deep learning-based tool named BioTMPy was developed to facilitate the search of relevant documents. BioTMPy is divided into separate modules that ease distinct processes of a document relevance pipeline. More precisely, modules to load datasets in different formats, convert them into distinct data structures, perform data analysis, and implement several deep learning models with their associated methods to perform hyperparameter optimization, cross validation, etc. Additionally, the package provides some examples on how to integrate all the modules together to perform a complete pipeline for document relevance.

To validate the developed pipelines, BioTMPy was later applied on a BioCreative's challenge from 2019. This challenge addressed the search of relevant documents for the topic of "mining protein interactions and mutations for precision medicine". With a comparison between different pre-trained embeddings, BioWordVec seemed to show on this data a slightly better performance over GloVe, "pubmed_pmc" and "pubmed_ncbi". Additionally, a model with a pre-trained BERT model (BioBERT) and a Bi-LSTM managed to surpass the best challenge's submission with a difference of 7.25% for average precision and 3.15% for the f1-score.

In addition, a web service was implemented to provide an effortless use of the developed model, allowing a user to order documents by relevance regarding the topic mentioned above. This means that, after gathering a corpus, one can use BioTMPy to develop a pipeline to retrieve the most relevant documents for a certain topic and consequently make it available to the public.

Session 3
7th May
11:30 h

Complete genome sequences of sixteen *Xanthomonas* spp. Strains assembled with short and long reads

Miguel Teixeira[1,2], Leonor Martins[1,2], Camila Fernandes[1,2,3], Cátia Chaves[1], Joana Pinto[1], Fernando Tavares[1,2,] Nuno A. Fonseca[1]

[1] CIBIO - Centro de Investigação em Biodiversidade e Recursos Genéticos, InBIO-Laboratório Associado, Universidade do Porto, Porto

[2] FCUP- Faculdade de Ciências, Departamento de Biologia, Universidade do Porto, Porto

[3] Unidade Estratégica de Investigação e Serviços de Sistemas Agrários e Florestais e Sanidade Vegetal, INIAV, Oeiras

Keywords: *Xanthomonas euroxanthea*, *Xanthomonas arboricola*, hybrid whole genome assembly.

Next generation sequencing technologies revolutionized microbial genomics, contributing to major advances in the last years. The use of a single technology is usually insufficient to guarantee simultaneously accurate and contiguous assemblies due to the trade-off between read length and accuracy. Short-reads have intrinsic issues with long repetitive regions, hence leading to fragmented assemblies. Long-reads usually have a higher error rate per base, thus preventing the detection of some genetic features.

To conciliate the advantages of both technologies we generated hybrid assemblies of 16 xanthomonad genomes with the Unicycler pipeline. The resulting assemblies, using the short and long-reads, are highly contiguous and recovered the expected genetic content with high confidence per base. It was observed that a misassemble of short-reads may compromise the overall contiguity. A second assembly pipeline was used to try to overcome this problem: long reads were assembled with Flye and the resulting assembly polished with the short reads using Pilon. This led to a highly contiguous assembly with high confidence per base.

The assembly of 16 whole genomes of walnut-associated strains (from *X. arboricola* and *X. euroxanthea* species) resulted in circular, high quality (Q>67) single. contig chromosomes and plasmids. These data disclose the whole genetic repertoire and genomic structure of the 16 strains, providing a foundation to access its shared evolutionary history and the emergence of genetic determinants of pathogenicity/virulence and host specificity.

Session 1
6th May
11:00 h

Deciphering the interaction of lactoferrin with V-ATPase towards a deeper understanding of its mechanisms of action

Cátia Santos-Pereira[1,2,3], Juliana F. Rocha[3], Henrique S. Fernandes[3], Lígia R. Rodrigues[2], Manuela Côrte-Real[1], Sérgio F. Sousa[3]

[1] Centre of Molecular and Environmental Biology (CBMA), Department of Biology, University of Minho, Braga, Portugal
[2] Centre of Biological Engineering (CEB), Department of Biological Engineering, University of Minho, Braga, Portugal
[3] UCIBIO@REQUIMTE, BioSIM, Departamento de Biomedicina, Faculdade de Medicina da Universidade do Porto, Porto, Portugal

Lactoferrin (Lf), a bioactive milk protein, exhibits strong anticancer and antifungal activities. The search for Lf targets and mechanisms of action is of utmost importance to enhance its effective applications. A common feature among Lf-treated cancer and fungal cells is the inhibition of a proton pump called V-ATPase. Lf-driven V-ATPase inhibition leads to cytosolic acidification, ultimately causing cell death of cancer and fungal cells. Given that a detailed elucidation of how Lf and V-ATPase interact is still missing, in this work we aimed to fill this gap by employing a multi-level computational approach. Molecular dynamics (MD) simulations of both proteins were performed to obtain a robust sampling of their conformational landscape, followed by clustering and protein-protein docking. Subsequently, MD simulations of the docked complexes and free binding energy calculations were carried out to evaluate the dynamic binding process and build the final ranking. This computational pipeline allowed the unraveling of the putative mechanism by which Lf inhibits V-ATPase and the identification of key binding residues that will certainly aid in the rational design of follow-up experimental studies, bridging in this way computational and experimental biochemistry.

Session 2
6th May
15:15 h

Dynamic genome-scale modelling of the *Saccharomyces* non-*cerevisiae* yeasts metabolism in wine fermentation

David Santos, D. Henriques, R. Minebois, A. Querol, E. Balsa-Canto

1. Departamento de Informática, Universidade do Minho, 4710-057, Braga, Portugal.
2. Grupo de Ingeniería de Bioprocesos, IIM-CSIC, Consejo Superior de Investigaciones Científicas, 6, 36208, Vigo, Espanha.
3. Department of Biotechnology, IATA-CSIC, Consejo Superior de Investigaciones Científicas, Paterna, 46980, Spain

The wine industry is facing challenging times due, mostly, to climate change and changing consumer demands. The urge to innovate stimulates R&D of new fermentation processes using non-conventional yeast species (*e.g.,* non-*cerevisiae Saccharomyces* species). Also, the use of low temperatures during fermentation promotes better aroma profiles. While recent research approached the physiology of diverse non-conventional yeast species, little is known about their metabolism in different environmental conditions.

In this work, a previously developed dynamic genome-scale model was adapted to study the metabolism of *Saccharomyces kudriavzevii* in wine fermentation at two temperatures, 25ºC and 12ºC. Adjustments included the addition of metabolic pathways and dynamic constraints. Goodness-of-fit of the model to measurements of the extracellular compounds was satisfactory, *i.e.,* the median values of R2 are 0.95 and 0.87 for 25ºC and 12ºC, respectively.

The model was then used to explore the differences in the dynamics of metabolism between temperatures. The most significant differences appeared in the stationary phase: 1) the strain produces more mevalonate and succinate at 25ºC, probably due to a late response to stress and the maintenance of redox balance via the GABA shunt, respectively, 2) erythritol flux is higher at 12ºC, probably due to the conditions of formation lasting longer and 3) the production of higher alcohols, mostly *de novo*, is higher at 12ºC, due to the longer viability of the cells.

The proposed model provided a comprehensive picture of the main steps occurring inside the cell during wine fermentation. Model predictions are consistent with experimental data and previous findings. The model also brought novel biological insights, such as the role of the GABA shunt or the production of mevalonate in the metabolism of *S. kudriavzevii*, worth being explored further.

Session 1
6th May
11:00 h

Gibbs Free Energy Variation Analysis of Breakpoint Regions from MtDNA Human Brain Deletions

João Carneiro[1], Brooke E. Hjelm[2,3], Filipe Pereira[4,5]

[1] Interdisciplinary Centre of Marine and Environmental Research (CIIMAR), University of Porto, Porto, Portugal
[2] Department of Psychiatry and Human Behavior, University of California-Irvine (UCI), Irvine, CA 92697, USA
[3] Department of Translational Genomics, Keck School of Medicine of USC, University of Southern California (USC), Los Angeles, CA 90033, USA
[4] IDENTIFICA genetic testing, Rua Simão Bolívar 259 3º Dir Tras, 4470-214 Maia, Portugal
[5] Departamento de Ciências da Vida - Universidade de Coimbra. Calçada Martim de Freitas, 3000-456 Coimbra, Portugal.
Corresponding author: João Carneiro - joaomiguelsov@gmail.com

In the last decade the importance of several non-B conformations that arise in both nuclear and mtDNA genomes in different biological processes has been shown. These structures were not only associated with a high number of diseases but also with regulatory and stability functions in cells. We evaluated the probability of occurrence of low free energy non-B conformation structures in breakpoint regions resulting from a mtDNA human brain deletion dataset. The analysis of 4489 human brain deletions breakpoints was done by running an automatic workflow to generate non-B conformation prediction of 100 nt breakpoint mtDNA regions. We used 3 control datasets, which included random, shuffle, and flanking datasets. The statistical results showed that there was not significant difference from the control datasets. The structure that presented more stability from all datasets was from the random sequences with a free energy value of -20.04 kcal/mol at trnN(gtt) (mtDNA position 5731). Nevertheless, we obtained an interesting result that showed several flanking regions presented very low free energy values (between -18 and -19.58 kcal/mol). The first non-B conformation with position at downstream gene (5' breakpoint dataset) and a very low free energy value (-18.71 kcal/mol) was at mtDNA position 5715. Further analysis will be done regarding the non-B conformations with lowest free energy values to understand their impact in relevant human brain biological processes in mtDNA.

INSaFLU: an open web-based bioinformatics suite for influenza and SARS-CoV-2 genome-based surveillance

Miguel Pinheiro[1], Ricardo J. Pais[2], Joana Isidro[2], João Paulo Gomes[2], Vítor Borges[2]

[1] Institute of Biomedicine-iBiMED, Department of Medical Sciences, University of Aveiro

[2] Bioinformatics Unit, Department of Infectious Diseases, National Institute of Health Dr. Ricardo Jorge, 1649-016 Lisbon, Portugal.

A new era of virus surveillance has already started based on the real-time monitoring of virus evolution at whole-genome scale. Although national and international health authorities have strongly recommended this technological transition, specially for influenza and SARS-CoV-2, the implementation of a routine and timely genomic surveillance, can be particularly challenging due to the lack of bioinformatics infrastructures and/or expertise to deal with primary next-generation sequencing (NGS) data.

We developed and implemented INSaFLU (“INSide the FLU”), which is an influenza and SARS-CoV-2 oriented bioinformatics free web-based suite that deals with primary NGS data (reads) towards the automatic generation of the output data that are actually the core first-line “genetic requests'' for effective and timely influenza and SARS-CoV-2 laboratory surveillance. By handling NGS data collected from any amplicon-based schema (making it applicable for other pathogens), INSaFLU enables any laboratory to perform multi-step software intensive analyses in a user-friendly manner without previous advanced training in bioinformatics.

INSaFLU gives access to user-restricted sample databases and project management, being a transparent and flexible tool specifically designed to automatically update project outputs as more samples are uploaded. Data integration is thus cumulative and scalable, fitting the need for a continuous epidemiological surveillance during the epidemics. Multiple outputs are provided in nomenclature-stable and standardized formats that can be explored in situ or through multiple compatible downstream applications for fine-tuned data analysis.

INSaFLU handles NGS data collected from distinct sequencing technologies (Illumina, Ion Torrent and Oxford Nanopore Technologies – ONT), with the possibility of construct analysis mixing technologies.

The bioinformatics pipeline consists of six core steps: (1) read quality analysis and improvement, (2) type and subtype identification, (3) variant detection and consensus generation, (4) coverage analysis, (5) alignment/phylogeny, (6) intra-host minor variant detection (and uncovering of putative mixed infections).

All the code is available in github (https://github.com/INSaFLU) with the possibility of a local docker installation (https://github.com/INSaFLU/docker). A detailed documentation and tutorial is also available (https://insaflu.readthedocs.io/en/latest/).

In summary, INSaFLU supplies public health laboratories and researchers with an open and user-friendly framework, potentiating a strengthened and timely multi-country WGS-based virus surveillance.

Session 1
6th May
11:00 h

MD simulations reveal that the parainfluenza FP forms oligomeric pore-like structures inside a membrane

M.F. Valério[1], D. Mendonça[2], J. Morais[2], C.H. Cruz[1], M.A.R.B. Castanho[2], S.S. Veiga[2], M.N. Melo[1], C.M. Soares[1], D. Lousa[1]

[1] ITQB NOVA, Instituto de Tecnologia Química e Biológica António Xavier, Universidade Nova de Lisboa;
[2] IMM, Instituto de Medicina Molecular João Lobo Antunes, Faculdade de Medicina, Universidade de Lisboa.

The parainfluenza virus (PIV) belongs to the large family of paramyxoviruses. Annually, these viruses contribute significantly to the global burden of disease in humans by infecting millions of individuals worldwide and leading to many deaths in areas with poor health care resources. During the infection process the virus must enter the host's cells by fusing its membrane with the host membrane. In the case of the parainfluenza virus, the cell entry process starts with the identification and attachment to target receptors, followed by proteolytic cleavage of the fusion glycoprotein (F) protein, exposing the fusion peptide (FP) region. The FP is responsible for binding and disturbing the target membrane3. It is believed to play a crucial role in the fusion process, however, the mechanism by which the parainfluenza FP peptide promotes membrane fusion is still unclear. To elucidate this matter, we performed coarse grain (CG) and atomistic molecular dynamics (MD) simulations, together with spectroscopic experiments of the parainfluenza FP in membranes. The combination of both these approaches led to the pinpointing of the most important residues for membrane fusion and the novel finding that this peptide, at high concentrations, induces formation of a pore-like structure. Our findings are a step further in the understanding of the membrane fusion process induced by the parainfluenza FP.

Session 1
6th May
11:00 h

New insights on SARS-Cov-2 M protein: Structure, Member Orientation and Mutations

Catarina Marques-Pereira[1,2**], Manuel N. Pires[1,2,3**], Nádia N. Pereira[1,2], Nícia Rosário-Ferreira[1,4], Irina S. Moreira[1,2,3*]

[1] University of Coimbra, Center for Neuroscience and Cell Biology, Coimbra, Portugal.
[2] University of Coimbra, Center for Innovative Biomedicine and Biotechnology, Coimbra, Portugal.
[3] University of Porto, Porto, Portugal
[4] Coimbra Chemistry Center, Department of Chemistry, University of Coimbra, Coimbra, Portugal.
[5] University of Coimbra, Department of Life Sciences, Calçada Martim de Freitas, Coimbra, Portugal.
* Email: irina.moreira@cnc.uc.pt
** Co-Authors

Severe Acute Respiratory Syndrome CoronaVirus-2 (SARS-CoV-2), a member of the Coronaviridae family, is the virus responsible for the COronaVIrus Disease (COVID-19). Since its first report on December 2019 in Wuhan, China over 108 M infection cases and 2.38 M deaths were reported worldwide (https://coronavirus.jhu.edu/map.html). SARS-CoV-2 virus comprises four major structural proteins: Spike (S), Envelope (E), Membrane (M) and Nucleocapsid (N), along with several accessory nonstructural proteins. The M protein is the most abundant protein in SARS-CoV-2 virion. Furthermore, this structural protein is responsible for a plethora of functions during the viral infection cycle along with a pivotal role in the interferon antagonism. Its high-level of conservation makes it a perfect target for drug discovery.

As there is a lack of experimentally obtained structures, the M protein structure and membrane orientation were predicted, revealing that the M protein has a small extracellular domain, a transmembrane domain involving three helices and a large intracellular domain. We analyzed the known M protein mutations using 433.425 SARS-CoV-2 genomes and 435.890 M protein sequences retrieved from the GISAID database (https://www.gisaid.org/). We identified 431 single mutations on the M protein: 27 mutations in the extracellular domain, 189 in the transmembrane domain and 215 in the intracellular domain. Mutant proteins stabilities and free binding energy differences were calculated to infer mutation impact in the overall protein structure as well as in their potential dimeric interfaces.

This work was funded by COMPETE 2020 - Operational Programme for Competitiveness and Internationalisation and portuguese national funds via FCT - Fundação para a Ciência e a Tecnologia, under projects DSAIPA/DS/0118/2020, UIDB/04539/2020 and LA/P/0058/2020. Catarina Marques-Pereira was also supported by FCT through Ph.D. scholarship 2020.07766.BD

Session 3
7th May
11:30 h

One bioinformatics pipeline for genomic characterisation of *Candida* sp. clinical strains

Carvalho MJ[1], Silva C[1], Guimarães R[1], Bezerra R[1], Pinheiro M[1], Santos MA[1], Moura G[1]

[1] Institute of Biomedicine- iBiMED, Department of Medical Sciences, University of Aveiro

Candida sp. are the most common cause of fungal infections. The increase of *Candida* antifungal resistance to the most widely used antifungals is alarming, given persistente infections are common and may be left untreated. Genomic surveillance of *Candida* isolates allows for the investigation of the biology, ecology, phylogenomics and

epidemiology of these pathogens. Importantly, the genomic characterisation of *Candida* isolates provides accurate species identification, monitoring of antifungal resistance, and the surveillance of the emergence of novel pathogens, which are crucial for appropriate diagnostic and treatment. On the other hand, as eukaryotic unicellular organisms, *Candida* sp. present specific genomic characteristics that distinguish them from the most commonly studied genomes,

i.e., bacteria and human. In this way, many of the open-source bioinformatic tools becomes inappropriate to tackle fungal genomes as needed, which hinders this type of analyses and delays their application to the clinical setting. Furthermore, Oxford Nanopore Technologies have been presented as an easy-to-use sequencing methodology in the clinical setting, but ONT-dedicated bioinformatic pipelines aiming to analyse and extract useful information from fungal genomes collected during infection are still lacking.

In this work, we have addressed this issue, and propose a bioinformatics pipeline to conduct isolates identification and variant analysis of fungal species with clinical relevance, starting with raw data from MinION sequencing of whole genomes of tem *Candida* sp. isolates from vaginal, oral and blood samples. Our results show the importance of genomic surveillance for the accurate identification of pathogens and understanding the adaptability of isolates to particular niches. Future work envolves Illumina sequencing of these genomes for hybrid assembly/read correction and improvement of our analysis pipeline.

CANCYL-POCI-01-0145-FEDER-031849 is supported by the Operational Program POCI, in its FEDER/FNR component, and the Foundation for Science and Technology, in its State Budget component. GenomePT POCI-01-0145-FEDER-022184; PBE POCIO1-0145-FEDER-031238; iBiMED UIDP/04501/2020-Programático

Session 1
6th May
11:00 h

Protein Engineering of Mini Therapeutic Proteins against SARS-CoV-2

Carlos Cruz[1], Carolina C. Buga[1], Susana Parreiras[1], Mariana Valério[1], Cláudio M. Soares[1], João B. Vicente[1] and Diana Lousa[1]

[1] ITQB NOVA, Instituto de Tecnologia Química e Biológica António Xavier, Universidade Nova de Lisboa, Oeiras, Portugal

The COVID-19 pandemic has become the most important health crisis in recent decades, with several critical outcomes, including the death of millions of people and severe economic consequences worldwide. In this scenario, protein engineering emerged as a powerful tool for the rapid development of potential therapeutic molecules to combat this disease. In this sense, this work focuses on the computational design of mini therapeutic proteins based on binder-receptor interactions knowledge and antigenic fragments. The goal was to create molecular platforms to i) block the interactions between the viral particle and its cell receptor in order to reduce the infection levels and ii) design antigenic proteins based on the fusion peptide and antigenic fragments from the Spike protein.

The SARS-CoV-2 Spike protein plays an important role in infection, attaching the virus at the host cell surface by binding to the receptor ACE2, thus initiating the fusion process. A reasonable assumption, therefore, is to prevent the Spike-ACE2 interaction by designing a competitive inhibitor of this interaction. For this purpose, the binding motif of ACE2 was transplanted onto scaffold proteins within the context of the Receptor-Binding Domain (RBD) using a motif-grafting method. The core, boundary and interface were then mutated on the presence of RBD to improve the binding affinity. The best designs were characterized by atomistic molecular dynamics (MD) simulations and new sequence solutions were proposed taking the simulations as a base. Another alternative to fight this virus lies in the development of proteins that carry small viral fragments capable of inducing immunity, acting as a vaccine. In this approach, the TOP7, an unnatural protein with high thermo-stability, was modified to include the SARS-CoV-2 fusion peptide, by means of grafting methods and massive mutagenic assays around the peptide. The protein dynamics, stability and folding were investigated through MD simulations and folding prediction methods, respectively. The designed molecules are being validated experimentally with promising results, showing that we have designed interesting lead molecules targeting SARS-CoV-2.

Session 2
6th May
15:15 h

Revising lipid chemical structures in genome-wide metabolic models with BOIMMG

Capela, J.[1], Liu, F.[2], Dias, O.[1]

[1] Centre of Biological Engineering, University of Minho, Campus de Gualtar, 4710–057 Braga, Portugal
[2] Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, USA

An important step in the reconstruction of Genome-Scale Metabolic (GSM) models is the integration of biochemical data. Such information is often incomplete or generic, lacking in completely defined chemical structures for several molecules, including lipids. The inumerous combinations of fatty acids in the side chains of lipids, hinder their storage in databases and integration into GSM models. Generic representations are commonly used to circumvent such limitation. However, lipid specificity is likely lost, and data integration problems arise, as several models contain lipids with completely defined structures and others with their generic versions. Such clash of versions is addressed by the Biochemical cOmplex data Integration in Metabolic Models at Genome-scale (BOIMMG). BOIMMG is an open-source framework that accelerates the swapping of different molecular versions (mainly lipids, structurally defined or not) in GSM models. Upon integration into a Neo4j graph database (http://neo4j.com/), lipid-specific data from LIPID MAPS Structure Database (LMSD), Swiss Lipids (SLM) and Model SEED were processed for biosynthetic contextualization within the curated pathways of MetaCyc. Several algorithms were developed to integrate this information in GSM models, afterwards.

Over 30 generic reactions were fully and 27 partially expanded, resulting in 557392 new reactions, in which 557252 were not integrated, nor listed in Model SEED. These reactions were inferred from the previously contextualized biosynthetic relationships between structurally defined compounds.

BOIMMG's information was applied to GSM models, tackling the conflict of molecules' versions. The whole glycerolipids and phospholipids' metabolic network within *E. coli* iJR904 model was expanded by our approach. The comparison between the altered model and one of its manually-expanded published iterations (iAF1260b), has shown that 53 and 38 more matching lipids and reactions, respectively, were found. Besides the new biochemical set, *BioISO*'s analysis demonstrated that biomass lipids were correctly produced, corroborating the correct expansion of the whole biosynthetic network. In conclusion, BOIMMG (available at https://boimmg.bio.di.uminho.pt/) can establish relevant relationships between complex macromolecules, within their biosynthetic context, and provide automated procedures for their integration into GSM models.

Session 3
7th May
11:30 h

Single-cell genomics in oligodendroglia: opening doors to understand multiple sclerosis

Ana Mendanha Falcão[1,6], David van Bruggen[1], Sueli Marques[1], Mandy Meijer[1], Sarah Jäkel[2], Eneritz Agirre[1], Samudyata[1], Elisa M. Floriddia[1], Darya Vanichkina[4,5], Charles ffrench-Constant[2], Anna Williams[2], André Ortlieb Guerreiro-Cacais[3], Gonçalo Castelo-Branco[1]

[1] Laboratory of Molecular Neurobiology, Department Medical Biochemistry and Biophysics, Karolinska Institutet, Stockholm, Sweden
[2] MRC Centre for Regenerative Medicine, Edinburgh bioQuarter, University of Edinburgh, Edinburgh, EH16 4UU, UK
[3] Department of Clinical Neuroscience (CNS), Karolinska Institutet, Stockholm, Sweden
[4] Gene and Stem Cell Therapy Program, Centenary Institute, University of Sydney, Australia
[5] Institute for Molecular Bioscience, University of Queensland, Australia
[6] Life and Health Sciences Research Institute, University of Minho, Braga, Portugal

Single-cell genomics have shaped our understanding of celular heterogeneity. Oligodendroglial cells, the myelinating cells of the central nervous system, were no exception. There are up to thirteen diferente oligodendroglial populations in the healthy central nervous system. What happens to these different populations in diseases such as multiple sclerosis (MS)? Do they react or are affected differently? To answer these questions we have performed single cell/nucleus transcriptomic analysis of oligodendroglial cells from both human postmortem MS tissue, and MS animal models. Excitingly, we found that unique oligodendroglial populations emerge in response to disease. Of notice, we uncovered a subset of oligodendrocytes and their progenitor populations expressing genes involved in antigen processing and presentation implying alternative functions of these cells in a disease context. Our results suggest that oligodendroglial cells are not passive targets but instead active immunomodulators in MS.

Session 1
6th May
11:00 h

Structure-based virtual screening, Molecular Dynamics and free energy calculations for the identification of novel inhibitors against biofilm formation by *C.violaceum*

Fábio G. Martins[1,2], André Melo[2], Sérgio F. Sousa[1]

[1] UCIBIO/REQUIMTE, BioSIM – Departamento de Medicina, Faculdade de Medicina da Universidade Do Porto, Alameda Prof. Hernâni Monteiro, 4200-319 Porto, Portugal
[2] LAQV/REQUIMTE, Departamento de Química e Bioquímica, Faculdade de Ciências da Universidade do Porto, Rua do Campo Alegre, 4169-007 Porto, Portugal

Biofilms are aggregates of microorganisms anchored to a surface and embedded in a self-produced matrix of extracellular polymeric substances. Bacteria within biofilm shave multiple advantages when compared to their planktonic counterparts. Biofilm infections have been recognized as a serious threat to our society. Quorum Sensing is an important process during biofilm maturation, in which cells communicate using autoinducer signals. Because quorum sensing has specific protein targets, it is possible to design inhibitors to block the formation of these structures.

Protein-Ligand molecular docking is a computational tool which predicts the binding pose and affinity of a ligand to a specific receptor or enzyme. During a virtual screening procedure, thousands of molecules are docked into a particular target and scored, giving an indication to which molecules are more probable to be active. MM/PBSA and MM/GBSA are used to estimate the free energy of the binding of small ligands to biological macro-molecules. These methods are based on molecular dynamics simulations of the receptor-ligand complex.

This work is focused in discovering new promising compounds against CviR, the quorum sensing receptor from *Chromobacterium violaceum*. Autodock 4, Autodock Vina, GOLD and LeDock were used in this work. The ability to discriminate the active molecules within a large database was optimized by screening a library containing known active molecules and decoys. The optimized protocol was then applied to a ZINC/FDA Approved database and to the Mu.Ta.Lig Virtual Chemotheca, which resulted in a list of promising compounds for further studies. Finally, Molecular dynamics simulations of the most promising molecules, in complex with CviR, were performed. Using the last 40 ns of simulation, MM/PBSA and MM/GBSA calculations were done in order to estimate the affinity of each molecule towards CviR. This study resulted in multiple promising compounds which in the future can be tested and validated experimentally.

Session 3
7th May
11:30 h

SynPred: Prediction of Drug Combination 1 Effects in Cancer using Full-Agreement Synergy Metrics and Deep Learning

António J. Preto[1,2], Pedro Matos-Filipe[1], Joana Mourão[1] and Irina S. Moreira[3,1*]

[1] University of Coimbra, Center for Neuroscience and Cell Biology, 3004-504 Coimbra, Portugal

[2] PhD Programme in Experimental Biology and Biomedicine, Institute for Interdisciplinary Research (IIIUC), University of Coimbra, Casa Costa Alemão, 3030-789 Coimbra, Portugal

3 University of Coimbra, Department of Life Sciences, Calçada Martim de Freitas, 3000-456 Coimbra, Portugal

* To whom correspondence should be addressed. Tel: (+351) 239 240 227; Email: irina.moreira@cnc.uc.pt

High-throughput screening technologies continues to produce large amounts of multiomics data from different populations and cell types for various diseases, such as cancer. However, analysis of such data encounters difficulties due to disease heterogeneity, further exacerbated by human biological complexity and genomic variability. Now is the time to redefine the approach to drug discovery, bringing an Artificial Intelligence (AI)-powered informational view that integrates the relevant scientific fields and explores new territories. Here, we show SynPred, an interdisciplinary approach that leverages specifically designed ensembles of AI-algorithms, links omics and biophysical traits to predict anticancer drug synergy. SynPred exhibits state-of-the-art performance metrics: accuracy – 0.85, precision – 0.77, recall – 0.75, AUROC – 0.82, and F1-score - 0.76 in an independent test set. Moreover, data interpretability was achieved by deploying the most current and robust feature importance approaches. A simple web-based application was constructed, allowing easy access by non-expert researchers.

Session 2
6th May
15:15 h

Towards an automatic cross-species comparative genomics portal

Jorge S. Oliveira[1], Pedro Pais[2,3], Miguel Antunes[2,3], Margarida Palma[2,3], Mónica Galocha[2,3], Cláudia P. Godinho[2,3], Inês V. Costa[2,3], Romeu Viana[2,3], Isabel Sá-Correia[2,3], Miguel C. Teixeira[2,3], Pedro T. Monteiro[1,4]

[1] INESC-ID, R. Alves Redol, 9, 1000-029 Lisbon, Portugal.

[2] Department of Bioengineering, Instituto Superior Técnico, Universidade de Lisboa, Av. Rovisco Pais, 1049-001 Lisbon, Portugal

[3] iBB-Institute for BioEngineering and Biosciences, Biological Sciences Research Group, Av. Rovisco Pais, 1049-001 Lisbon, Portugal

[4] Department of Computer Science and Engineering, Instituto Superior Técnico, Universidade de Lisboa, Av. Rovisco Pais, 1049-001 Lisbon, Portugal

The recently created YEASTRACT+ portal (http://yeastract-plus.org/), is comprised of three distinct yet interconnected databases: Yeastract (*S. cerevisiae*), PathoYeastract (Pathogenic Yeast) and NCYeastract (Non-conventional Yeast). It currently holds information of ten yeast species, gathered for over 15 years, permitting the development of improved tools for the prediction of gene and genomic regulation based on orthologous regulatory associations described for other yeast species, as well as visualization tools for cross-species transcription regulatory networks [Monteiro, NAR 48-D1:642–649 2020].

Numerous genomes are sequenced and made available to the community through the NCBI portal. However, unlike what happens for gene function annotation, annotation of promoter sequences and the underlying prediction of regulatory associations is mostly unavailable, severely limiting the ability to interpret genome sequences in a functional genomics perspective.

Here, we present a semi-automatic approach where one can download a genome of interest from NCBI in the GenBank Flat File (.gbff) format and, with a minimum set of commands have: all the information parsed, organized, and made available through the platform web interface.

This approach motivated the creation of an additional database -- CommunityYeastract. Genomes deposited in this database can thus be compared to any genome of reference, from the YEASTRACT+ portal, in search of homologous genes, shared regulatory elements and predicted transcription associations. Alternatively, non-yeast communities can install the platform independently, insert one or several genomes, without any constraints, and take advantage of the data visualization, analysis and comparison tools. All the code and step-by-step instructions to download, pre-process, and load a given genome of interest in during the current study are available in the GitLab repository, https://gitlab.com/oliveira.jorge.88/web

Along with the interconnectivity of YEASTRACT+ and the new Community section, a set of Cross Species tools have been developed: to compare and find unique TF binding sites in homologous genes, by comparing their promoters; to view regulatory networks for the same transcription factors and genes across multiple species; to directly compare a network between two species in order to find unique and common regulations; and to rank transcription factors for a given set of regulated homologous genes.

Funding: LISBOA-01-0145-FEDER-022231 - BioData.pt Research Infrastructure, EC H2020 grant 676559 ELIXIR-EXCELERATE, FCT project PTDC/BII-BIO/28216/2017 and Plurianual UID/CEC/50021/2020.

Session 1
6th May
17:00 h

3D Prediction of Non-B DNA Conformations Associated with MtDNA Genomic Instability

André F. Pina[1], Sérgio F. Sousa[1], João Carneiro[2]

[1] UCIBIO@REQUIMTE, BioSIM, Departamento de Biomedicina, Faculdade de Medicina, Universidade do Porto, Portugal
[2] Interdisciplinary Centre of Marine and Environmental Research (CIIMAR), University of Porto, Porto, Portugal
Corresponding author: João Carneiro - joaomiguelsov@gmail.com

Non-B DNA conformations are molecules that do not follow the canonical double helix DNA structure. This type of molecules, which can assume both simple (e.g. hairpins) and complex (e.g. G-quadruplexes) conformations, are highly related with mutagenetic instability in both nuclear and mitochondrial DNA (mtDNA). We have previously detected a cloverleaf-like non-B conformation (Structure A) predicted for a 93-nt (nucleotide) stretch of the mtDNA control region 5′-peripheral domain. The Structure A occurs in a hot spot for the 3′ end of human mtDNA deletions. To better characterize the Structure A we predicted the 3D conformation using state-of-art algorithms and methods. The 3D model was built in RNAComposer using as input the UNAFold predictions. We converted the RNA to DNA using the Amber software package (xleap). The molecular dynamics simulations were made considering three control structures (93 nt Structure A mono-nucleotide shuffle sequence, 93 nt sequences with highest and lowest folding potential in mtDNA). MD simulations were performed with the OL15 force field, developed for atomistic simulations of nucleic acids. Initial simulations were run for 10 ns at 310 K and 1 atm, in a box of waters, under periodic boundary conditions. The results illustrate the stability and level of base pairing of the different structures and open the door for an atomic level understanding of non-B DNA stability.

Session 2
7th May
10:15 h

Bioinformatic analysis of promoters from *Ashbya gossypii*

Pedro Montenegro-Silva[1], Lucília Domingues[1] and Tatiana Aguiar[1]

[1] CEB-Centre of Biological Engineering, University of Minho, 4710-057 Braga, Portugal

Ashbya gossypii is a filamentous fungus that produces large quantities of riboflavin, a living factory that generates nearly half of this vitamin's world supply. The fungus can grow on industrial wastes and can produce other value-added compounds including recombinant proteins, single cell oil, nucleosides, folic acid and organoleptic compounds. At the genome level, 95% of the genes have homologues in *Saccharomyces cerevisiae* (90% in syntenic positions). Many molecular tools used in yeast are functional in both organisms, including the autonomously replicating sequences.

Fine-tuning heterologous metabolic pathways requires the availability of promoters with a wide range of activity. Endogenous promoters provide the main regulatory elements for gene expression control, however, there is a limited range of well characterised promoters available for metabolic engineering of *A. gossypii*. Repeated copies of promoters for genome editing purposes increases the probability of homologous recombination and causes strain instability, what is disadvantageous when extensive pathway engineering is performed.

The identification of transcription factor (TF) binding sites (TFBSs) in promoters is important for rational design of regulatory crosstalk in metabolic engineering strategies. TFs usually bind to degenerated sequences that can be identified using bioinformatics tools. The MEME algorithm can be used to "mine" DNA motifs from upstream sequences of co-expressed or co-regulated genes. The obtained motifs can further be compared to TFBSs in databases for *S. cerevisiae* (*i.e.* JASPAR, TRANSFAC, YEASTRACT). This procedure allowed the identification of 4 DNA motifs, 8 matching putative TFs, and TATA-box as important elements for high level gene expression.

Thus, the bioinformatic approach used in this study, in intergenic regions from *A. gossypii*, can be also performed in sequences of other organisms with the aim of identifying potential candidate motifs for subsequent experimental characterization and allowing future construction of hybrid semi-synthetic promoters.

Supported by FCT, Compete2020 and Norte2020. Strategic funding of UID/BIO/04469/2019, Project EcoBioInks4SmartTextiles (PTDC/CTM-TEX/30298/2017–POCI-01-0145-FEDER-030298).

Session 1
6th May
17:00 h

Bioinformatic Applications for the Development of Plastic Degradation Enzymes using QM/MM

Jorge M. Cunha, Rita P. Magalhães, Henrique S. Fernandes, Sérgio F. Sousa

UCIBIO/REQUIMTE, BioSIM - Departamento de Biomedicina, Faculdade de Medicina, Universidade do Porto, Alameda Professor Hernâni Monteiro, 4200-319 Porto, Portugal

The production of plastic has been dramatically increasing these past years, reaching values as high as 350 million tons annually. The excess when not treated properly, reaches the soils and oceans, causing its accumulation and bringing to rise major consequences to all forms of life. Thus, strategies for plastic degradation have been developed. Although the chemical strategies used in plastic degradation industry can degrade the plastic, the aftermath is not environmentally friendly, causing problems such as air pollution. For that reason, strategies like using biocatalysts are being developed. However, the understanding of how enzymes work remains, sometimes, unexplained at the molecular level.

Poly(ethylene terephthalate) (PET) is one of the most rigid polymers and is used to produce bottles, cleaning products, fibers in the textile industry, etc. Currently, only a few reported enzymes can biodegradate the polymer, however the enzymatic activity is very low when compared to industry levels. PETase and MHETase, two promising enzymes for PET biodegradation, from the bacterium *Ideonella sakaiensis* were recently discovered in Japan near plastic recycling locations, offering appealing prospects for the development of improved biocatalyists.

QM/MM methods combine both quantum chemical treatment (QM), applied in the electronically important region allowing modelling of chemical reactions, where the active site is comprised, which is directly involved in the chemical reaction and mechanical treatment (MM), applied in the region that surrounds the electronically important region, encompassing the rest of the system.

This work reports the application of QM/MM methods in the study of the catalytic mechanisms of PETase and MHETase aiming for future enhancement of the PET degradation rate by site-directed mutagenesis.

Session 2
7th May
10:15 h

Bioinformatic approaches to address new perspectives on genotype-phenotype associations for complex diseases

Daniel Martins[1,2], Conceição Egas[2], Joel Arrais[1]

[1]. CISUC - Centre for Informatics and Systems of the University of Coimbra. Polo II, Pinhal de Marrocos, 3030-290 Coimbra, Portugal.

[2]. CNC - Center for Neuroscience and Cell Biology of the University of Coimbra (UCBiotech). Parque Tecnológico de Cantanhede, Núcleo 04, Lote 8, 3060-197 Cantanhede, Portugal.

Complex diseases like Type 2 diabetes (T2D) have been a matter of interest in biomedical research. Although the number of disease-associated variants has increased, its combined genetic risk has not surpassed the value of 15%.

Despite the involvement of several pathways and biological networks on T2D pathophysiology, most genotype-phenotype association studies on the disease have been conducted on GWAS data assuming biological independence between genetic variant effects. Polygenic effects are, therefore, mostly addressed from a theoretical point-of-view.

On this scope, various models have been proposed to explain how genetic variants contribute to a given phenotype. The discussion stays relevant on the present day, as one of the most recent, the omnigenic model, was proposed in 2017 by Boyle *et al*, arising a new perspective on the ongoing debate.

As biomedical data is produced at an exponential rate and the knowledge on genomics and proteomics grows accordingly, bioinformatics might play an essential role in testing and providing new insights to the theoretical models on genotype-phenotype associations. Either by integrating information from different sources into new metrics or improving the representation and interpretation of gene interactions and combined effects. Machine and Deep Learning approaches are ideal for these tasks due to its ability to learn the correlations on data and explore its representation under different abstractions.

Here, we aim to present an overview of the current state of the art on complex diseases research, specifically on Type 2 Diabetes, and possible contributions of bioinformatic approaches for the analysis and exploration of gene-gene interactions and their influence on the manifestation of complex traits.

This work is financed by FCT - Fundação para a Ciência e Tecnologia
Scholarship reference: SFRH/BD/146094/2019

FCT Fundação para a Ciência e a Tecnologia CENTRO 2020 PORTUGAL 2020 UNIÃO EUROPEIA Fundo Europeu de Desenvolvimento Regional

Session 2
7th May
10:15 h

Comparison of bioinformatics tools to predict the presence of prophages in *Helicobacter pylori* genomes

Rute Ferreira[1,3], Cláudia Sousa[1], Eva Presa[1], Diana P. Pires[1], Mónica Oleastro[2], Joana Azeredo[1], Céu Figueiredo[3,4,5], Luís D. R. Melo[1]

[1] CEB –Centre of Biological Engineering, University of Minho, Braga, Portugal
[2] Department of Infectious Diseases, National Institute of Health Doctor Ricardo Jorge (INSA), Lisbon, Portugal;
[3] i3S –Institute for Research & Innovation in Health, Porto, Portugal;
[4] Ipatimup – Institute of Molecular Pathology and Immunology of the University of Porto;
[5] Department of Pathology, Faculty of Medicine, University of Porto, Porto, Portugal

Bacterio(phages) are specific viruses for bacteria, being their natural enemies. When the genome of a phage is integrated into the host bacterial genome, it is named prophage. These are a latent form of phages, in which the viral genes can increase the virulence and/or fitness characteristics of the host. This life cycle - lysogenic - does not cause the bacterial cell to rupture. Prophages have already been identified in most pathogenic bacteria, providing them better chances of survival. In the case of *Helicobacter pylori*, a human gastric pathogen that causes, among others, chronic gastritis, peptic ulcers, and adenocarcinoma, the presence of important prophage genes in their genomes has already been identified. In our work, a total of 109 complete genomes of human isolates of *H. pylori* and plasmids, deposited in NCBI archives, between November 5, 2015, and February 21, 2020, and 19 complete genomes of Portuguese clinical isolates, were screened, regarding the presence of prophages. For that, two of the most widely used web servers for identifying putative prophages in bacterial genomes were used: Phaster and Prophage Hunter. With the use of Phaster, 78 prophage sequences were identified, 6 of which were intact (7.7 %). Regarding Prophage Hunter, 199 prophages were identified, in a total of 17 active (8.5 %). The differences observed in the number of prophages identified by each tool is probably due to variances in the identification methods that each tool uses, as already reported. However, the intact sequences identified in Phaster were also predicted, in the same strains, in Prophage Hunter. These results suggest a high probability of these strains having inducible sequences of prophages in their genomes. The use of web servers for the rapid identification and annotation of prophage sequences in bacterial genomes and plasmids has been growing, helping to direct laboratory experiments more easily. In this work, we observed some differences in the results between the two tools used, concluding that new prophage prediction tools using Machine-Learning are required to predict more accurately this important viral sequences.

Session 2
7th May
10:15 h

De novo assembly and annotation of the *Candida cylindracea* genome: a pipeline for rare organisms

Carvalho MJ[1], Pinheiro M[1], Santos MA[1], Moura G[1]

[1] Institute of Biomedicine- iBiMED, Department of Medical Sciences, University of Aveiro

Candida cylindracea is a biotechnologically long-used species, especially due its richness in low-specificity lipases that have been extensively used for bio detergent production. Additionally, CUG codons (leucine codons) are 100% translated as serine by *C. cylindracea*, representing a crucial evolutionary point in yeast phylogenetic history.

As eukaryotic unicellular organisms, Candida sp. present specific genomic characteristics that distinguish them from the most commonly studied genomes (bacterial and human), rendering the majority of open-source bioinformatics tools inappropriate for the study of fungal genomes. In the case of rare pathogens and rare organisms such as *C. cylindracea*, for which no genome sequence reference is available, the lack of such bioinformatic tools further hinder the study of such organisms and, in the case of pathogens, resolving outbreaks. Oxford Nanopore Technologies have been presented as an easy-to-use sequencing methodology, providing long reads spanning long genetic regions, which facilitate the assembly of genomes. However, ONT-dedicated bioinformatic pipelines aiming to analyse and extract useful information from fungal genomes are also lacking.

We started a bioinformatics pipeline to assemble *de novo* and annotate previously unknown fungal genomes, using *Candida cylindracea*. ONT sequencing resulted in >1M reads, with mean read length of >11Kbp and mean read quality (Phred score) of 10.3. Preliminary results assembled the >10Mbp genome in 17 contigs ≥10000bp using ONT only reads (Flye) whereas hybrid assembly using ONT and Illumina reads (SPAdes) produced 38 scaffolds ≥10000bp. Distinct polishing steps using DNA and RNA sequencing data have been tested to improve our analysis pipeline for full annotation of this fungus.

CANCYL-POCI-01-0145-FEDER-031849 is supported by the Operational Program POCI, in its FEDER/FNR component, and the Foundation for Science and Technology, in its State Budget component. GenomePT POCI-01-0145-FEDER-022184; PBE POCIO1-0145-FEDER-031238; iBiMED UIDP/04501/2020-Programático

Session 1
6th May
17:00 h

Decoding Partner Specificity in Opioid Receptor Family

Carlos A. V. Barreto[1,2,3], Salete J. Baptista[2,4], Daniel Silvério[1,2,3], António J. Preto[1,2,3], Rita Melo[2,4], Irina S. Moreira[5,1,2]

[1] University of Coimbra, Center for Innovative Biomedicine and Biotechnology, Coimbra, 3004-504, Portugal.
[2] University of Coimbra, Center for Neuroscience and Cell Biology, Coimbra, 3004-504, Portugal.
[3] University of Coimbra, Institute for Interdisciplinary Research, Coimbra, 3030-789, Portugal.
[4] Centro de Ciências e Tecnologias Nucleares, Instituto Superior Técnico, Universidade de Lisboa, Bobadela, Portugal.
[5] University of Coimbra, Department of Life Sciences, Calçada Martim de Freitas, Coimbra, 3000-456, Portugal.

Opioid receptor (OR) family is involved in several physiological processes, with a particular role in analgesia. However, the prolonged use of opioids (permanent stimulation) can lead to the increase of organism tolerance to these compounds, reducing their clinic effects, ultimately triggering opioid use disorder. This is the basis of the opioid crisis, a huge socio-economic issue worldwide, which is directly related to drug abuse.

ORs function is strictly dependent on the specific signaling pathways triggered by the coupling of ORs with the different intracellular partners. In order to gain further insights into physiological and pharmacological properties of these important drug targets, an extensive characterization of all members of ORs family (μ (MOR), δ (DOR), κ (KOR), nociceptin (NOP)) and their corresponding binding partners (ARRs: ARR2, ARR3; G-protein: Gi1, Gi2, Gi3, Go, Gob, Gz, Gq, G11, G12, G14, G15, Gs(sh), Gs(lo)) was performed. A multi-step approach including homology modelling, docking and molecular dynamic simulations was applied, after which a detailed description of the interaction interfaces was carried out. Overall, 68 complexes were built and both structural and dynamic analysis were performed. All data was compiled in a freely available website.

This approach represents a novel and exciting big data analysis of OR-partners interface determinants and establishes a further step into the understanding of OR family functional specificity.

This work was funded by COMPETE 2020 - Operational Programme for Competitiveness and Internationalisation and portuguese national funds via FCT - Fundação para a Ciência e a Tecnologia, under projects DSAIPA/DS/0118/2020, PTDC/QUI-OUT/32243/2017 and UIDB/04539/2020. Carlos A.V. Barreto was also supported by FCT through Ph.D. scholarship SFRH/BD/145457/2019

Session 1
6th May
17:00 h

Development of a Hybrid Metabolic Model for the Optimization of Aromatic Amino Acids in *E. coli.*

Leslie Avendaño-Montoya[1], Isabel Rocha[2], Sónia Carneiro[3]

[1,2] Instituto de Tecnologia Química e Biológica António Xavier
[1,2] Universidade NOVA de Lisboa
[3] SilicoLife

Mathematical models are used for data analysis, simulation and optimization of metabolic systems. Currently, one of the most common approaches is the constraint-based modelling based on mass balance constraints allowing a high coverage of the metabolic pathways. However, this methodology may generate inaccurate phenotype predictions as it is based on modelling assumptions regarding cellular behavior or product formation that might not fit *in vivo* observations. Other models, as kinetic models that do not need this type of assumptions, need experimental data and the definition of rate laws, making it difficult to have a large model covering most of the metabolic activities existing in a biological system. To improve the predictions, we propose a hybrid model that integrates additional biological information from GECKO, kinetic and regulatory models. This model, assembled for *E. coli,* was evaluated under different environmental conditions and using different regulatory rules. As a case-study, the model will be applied for predicting bottlenecks and possible genetic modifications that increase the aromatic amino acid production without affecting cellular viability. As the pathway involved, the shikimate pathway, is highly regulated, this approach allows evaluating different strategies that include also regulatory factors. Quasi-steady-state conditions were assumed for modelling fast reactions and a time delay parameter was incorporated for the slow reactions. Preliminary results show that phenotype prediction improve when the GECKO model is constraint with kinetic fluxes in steady state and regulatory rules. Finally, since the model does not include kinetic expressions for all metabolic reactions, an exhaustive list of kinetic parameters is not necessary. Furthermore, the level of coverage and detail is higher than for the stoichiometric models. For this reason, the hybrid model allows to improve flux predictions.

Session 1
6th May
17:00 h

Development of a Structural Database of Insecticide Targets

Maria F. Araújo[1,2*], Tatiana F. Vieira[1], Elisabete M.S. Castanheira Coutinho[2] and Sérgio F. Sousa[1]

[1] UCIBIO/REQUIMTE, BioSIM Departamento de Biomedicina, Faculdade de Medicina, Universidade do Porto, Alameda Professor Hernâni Monteiro, 4200-319 Porto, Portugal
[2] Departamento de Física, Escola de Ciências, Universidade do Minho, Rua da Universidade, 4710-057 Braga, Portugal
*mffaraujo@live.com.pt

To meet the needs of an exponentially growing population in terms of sustainable food production, an increase in the use of pesticides is inevitable to ensure a greater production and a safe food supply. However, despite their beneficial role, some of these agrochemicals have been associated to dangerous characteristics, including carcinogenicity, teratogenicity, high and acute residual toxicity, interference with the hormonal and reproductive systems of mammals and long environmental persistence. Thus, the development of alternative pesticides that are eco-friendly, safe to humans and non-target organisms and that can circumvent the evolution of resistance has been an important topic of research in recent years. This implies understanding the mode of action of conventional pesticides and knowing their targets.

In this work, we have developed a database gathering atomic level information on the protein targets directly associated to insecticide action. This database can be a valuable tool for those who want to study the mode of action of known pesticides, and the targets involved at a molecular level, or to develop or assess new molecule entities with possible insecticide action.

Currently, this database contains X-ray crystallographic data for 307 protein targets. It is organized in a scheme that associates the mode of action in broad categories based on the affected physiological functions, and considers parameters such as the source organism, the experimental method used, the existence of mutations on the target, and the presence or absence of a crystallographic ligand.

Session 2
7th May
10:15 h

Development of computational tools for the analysis of 2D-nuclear magnetic resonance data

Bruno Pereira[1], Marcelo Maraschin[2] and Miguel Rocha[1]

[1] Centre of Biological Engineering, University of Minho, Braga, Portugal

[2] School of Agricultural Sciences, Federal University of Santa Catarina, Santa Catarina, Brazil

Metabolomics is one of the omics' sciences that has been gaining a lot of interest due to its potential on correlating an organism's biochemical activity and its phenotype. The main techniques that collect data are based on mass spectrometry and nuclear magnetic resonance (NMR) spectroscopy. The last one has the advantage of analysing a sample *in vivo* without damaging it and while its sensitivity is pointed out as a disadvantage, multidimensional NMR delivers a solution to this issue. It adds layers of information, generating new data that requires advanced bioinformatics methods to extract biological meaning. The need to establish an integrated framework has become imperative due to different approaches that multidimensional NMR has, to tackle reproducibility issues across research groups.

In recent work from the host group, *specmine*, an R package for metabolomics and spectral data analysis/mining, has been developed and improved to wrap and deliver key metabolomic methods that allow a researcher to perform a complete analysis. Tools integrated in *specmine* were developed to read, visualize, and analyse two-dimensional (2D) NMR. A new *specmine* structure was created for this type of data, easing interpretation and data visualization. In terms of visualization a novel approach towards three-dimensional environments enables users to interact with their data. The selection of which samples to plot, when the user does not specify an input, is based on a signal-to-noise ratio scale and a method to perform peak detection on 2D NMR based on local maximum search was implemented to obtain a data structure that best benefits from *specmine*'s functionalities. These include pre-processing, univariate, and multivariate analysis as well as machine learning and feature selection methods.

The 2D NMR functions were validated using experimental data from two scientific papers, available on metabolomic databases. These data originated two case studies from different NMR sources, Bruker and Varian, which reinforces *specmine*'s flexibility. The case studies were carried out using mainly *specmine* and other packages for specific processing steps. A pipeline to analyse 2D NMR was added to *specmine*, in a form of a vignette, to provide a guideline for the newly developed functionalities.

Session 2
7th May
10:15 h

Evolving meaning for supervised learning in complex biomedical domains using knowledge graphs

Rita T. Sousa, Sara Silva, Catia Pesquita

LASIGE, Faculdade de Ciências da Universidade Lisboa

In recent years, the explosion in complexity and heterogeneity of data has motivated a new paradigm, where millions of semantically-described entities are available in knowledge graphs (KGs). Different KGs have been exploited in a wide variety of data mining and machine learning tasks, namely the prediction of specific relations between entities that correspond to KG instances but whose relationship is not encoded in the graph. This scenario has various bioinformatics applications, such as predicting protein-protein interactions, drug-drug interactions or gene-disease associations.

Many of the existing KG-based approaches for machine learning use KGs for generating static semantic representations, which are then used as features. These semantic representations can be considered static since they consider the full graph, blind to the fact that unnecessary information for representations can introduce noise. In applications where the target is encoded in the KG, this problem is mitigated. However, when the classification targets are not a part of the KG, representations cannot be trained on the targets. The problem is exacerbated in complex domains, such as the biomedical, where KGs represent multiple views (or semantic aspects) over the underlying data, some of which may be less relevant to train the model. This brings up the challenge of tailoring the semantic representation of the KGs entities to a specific goal.

Our research aims to address this challenge by developing novel machine learning-based approaches to learn suitable semantic representations of data objects extracted from KGs to support specific supervised learning tasks in bioinformatics applications. These novel approaches are anchored in a framework that integrates semantic representation and machine learning approaches, allowing a comparative evaluation of different combinations. This framework was successfully applied to protein-protein interaction prediction with significant improvements over manually defined static semantic representations, and to learn similarity functions adapted to different biological perspectives. Future work will include application to gene-disease association prediction.

This work was supported by FCT through LASIGE Research Unit (UIDB/00408/2020 and UIDP/00408/2020), the PhD grant (SFRH/BD/145377/2019) and the projects SMILAX(PTDC/EEI-ESS/4633/2014), BINDER(PTDC/CCI-INF/29168/2017) and PREDICT(PTDC/CCI-CIF/29877/2017). It was also partially supported by the KATY project funded by Horizon 2020 under grant agreement No 101017453.

Session 2
7th May
10:15 h

Exploring bioinformatics tools and databases to decipher the proteome of *Candida glabrata* biofilm matrix

Bruna Gonçalves[1] and Sónia Silva[1*]

[1]Centre of Biological Engineering, University of Minho, 4700-057, Braga, Portugal
d6643@ceb.uminho.pt and soniasilva@deb.uminho.pt; *corresponding author

Candida glabrata is a clinically relevant human pathogen with ability to form high recalcitrant biofilms, which produce an extracellular matrix suggested to have structural, virulent and protective roles. Thus, elucidation of matrix components, their function and regulation, is crucial to disclose matrix role in *C. glabrata* pathogenesis. As such, this study aimed to reveal, the matrix proteome of *C. glabrata* biofilms and to characterize it exploring bioinformatics tools. For that, extracted *C. glabrata* matrix proteins were analyzed through LC-MS/MS and identified using UniProt database. The functional distribution of the matrix proteins found was assessed using FungiFun tool and FunCat database. This analysis revealed an enrichment of proteins involved in carbohydrate-metabolism, which have a potential role in the delivery of carbohydrates into the matrix. Virulence-related functions were also found to be enriched among matrix proteins. Additionally, the predictive secretory nature of proteins was analysed using the Fungal Secretome Database and the Fungal Secretome KnowledgeBase. These analyses revealed that many matrix proteins have unconventional secretory pathways. Furthermore, orthologous proteins, identified with PathoYeastract database, were searched in *Candida* Genome Database using the keywords "extracellular region" and "biofilm matrix". High overlap between *C. glabrata* matrix proteins and those secreted by other *Candida* spp, especially those of *Candida albicans* biofilm matrix was confirmed. Finally, using PathoYeastract platform, Pdr1 was indicated to be a potential regulator of matrix proteome and STRING analysis revealed high molecular interaction, either direct or indirect, between Pdr1 target-proteins. This study provides a unique resource for further functional investigation of matrix proteins, contributing to the identification of potential targets for the development of new therapies against *C. glabrata* biofilms.

Session 2
7th May
10:15 h

Genetic risk for COVID-19 outcomes in COPD and differences among worldwide populations

Marçalo, R.[1,2], Neto, S.[1], Pinheiro, M.[1], Rodrigues, A.J.[3], Sousa, N.[3], Santos, M.A.S.[1], Simão, P.[4], Valente, C.[5], Andrade, L.[5], Marques, A.[2], and Moura, G.R.[1]

[1] Genome Medicina laboratory, Institute of Biomedicine.iBiMED, Department of Medical Sciences, University of Aveiro, Aveiro (Portugal).
[2] Lab 3R - Respiratory Research and Rehabilitation Laboratory, School of Health Sciences, University of Aveiro (ESSUA), Aveiro (Portugal).
[3] Life and Health Sciences Research Institute (ICVS), School of Medicine, University of Minho, Braga (Portugal).
[4] Pulmonology Department, Unidade Local de Saúde de Matosinhos, Porto (Portugal).
[5] Pulmonology Department, Centro Hospitalar do Baixo Vouga, Aveiro (Portugal).

People with chronic obstructive pulmonary disease (COPD) constitute one of COVID-19 risk groups for poor prognosis upon infection. Variability in predisposition and clinical response to COVID-19 exist but our understanding of these factors in the COPD population is limited. This study explored the genetic background as a possible answer to COVID-19 infection response heterogeneity, either for the poor prognosis in people with COPD or across healthy worldwide populations.

Significant SNPs (susceptibility: rs286914/rs12329760; severity: rs657152/rs11385942) were selected from the literature and their allelic frequencies used to calculate the probability of having multiple risk alleles in both our COPD cohort and each worldwide population. A polygenic risk analysis was conducted in the COPD cohort for the two mentioned phenotypes and for hospitalization and survival to COVID-19 infection.

No differences in genetic risk for COVID-19 susceptibility, hospitalization, severity or survival were found between people with COPD and the control group (all p-values>0.01), either considering risk alleles individually, allelic combinations or polygenic risk scores. Alternatively, all populations, even those with European ancestry (Portuguese/Spanish/Italian), showed significant differences from the European population in genetic risk for COVID-19 susceptibility and severity (all p-values<0.0001).

Our results indicated a low genetic contribution for COVID-19 infection predisposition or worse outcomes in people with COPD. Also, our study unveiled a high genetic heterogeneity across major world populations for the same alleles, even within European subpopulations.

This work was supported by FEDER (Fundo Europeu de Desenvolvimento Regional) funds through the COMPETE 2020, Operational Programme for Competitiveness and Internationalization (POCI) (POCI-01-0145-FEDER-028806;POCI-01-0145-FEDER-016428), CENTRO 2020 (CENTRO–46-2016-02) and by Portuguese national funds via Fundação para a Ciência e a Tecnologia, I.P. (FCT) (PTDC/DTP-PIC/2284/2014;PTDC/SAU-SER/28806/2017;PTDC/BIA-MIC/31849/2017). The iBiMED is supported by FCT funds under UIDP/04501/2020.

Session 1
6th May
17:00 h

Identification of Biofilm Formation Inhibitors in *Pseudomonas aeruginosa* through Virtual Screening, Molecular Dynamics, and Free Energy Calculations

Rita P. Magalhães[1*], Tatiana F. Vieira[1], André Melo[2] and Sérgio F. Sousa[1]

[1] UCIBIO/REQUIMTE, BioSIM – Departamento de Medicina, Faculdade de Medicina da Universidade Do Porto, Alameda Prof. Hernâni Monteiro, 4200-319 Porto, Portugal
[2] LAQV/REQUIMTE, Departamento de Química e Bioquímica, Faculdade de Ciências da Universidade do Porto, Rua do Campo Alegre, 4169-007 Porto, Portugal

Some bacteria form biofilms - highly organized communities attached to a surface and enclosed in a self-produced matrix. These structures are highly resistant to antibiotics and host immune response and affect both human tissues and medical devices. *P.aeruginosa* is a highly pathogenic and resistant gram-negative bacteria. The development of potent inhibitors against its mechanisms of biofilm formation is a promising therapeutic strategy to combat *P.a* related infection.

Virtual Screening (VS) is the application of molecular docking to large databases of compounds. Molecular Dynamics (MD) is a computational technique that simulates a flexible molecular system as a function of time. Free Energy Calculations (MM/P(G)BSA) allow for the calculation of binding free energy values between protein and ligand to predict their inhibitory potential.

A protocol combining the three computational techniques described was developed and employed to identify potent new inhibitors against biofilm formation in *Pseudomonas aeruginosa*. A total of 294,498 compounds were screened against the *LasR* receptor. 23 top scoring ligands were further investigated through MD and MMP(G)BSA calculations. We suggest 5 compounds as highly promising inhibitors to be tested experimentally and used as scaffold for further drug design campaigns.

This work was supported by the Applied Molecular Biosciences Unit - UCIBIO which is financed by national funds from FCT (UIDP/04378/2020 and UIDB/04378/2020).

Session 1
6th May
17:00 h

Implementing a webserver for managing and detecting viral fusion proteins

Pedro Moreira[1], Miguel Rocha[1], Diana Lousa[2]

[1] Escola de Engenharia da Universidade do Minho, Braga, Portugal
[2] ITQB NOVA, Instituto de Tecnologia Química e Biológica António Xavier, Universidade Nova de Lisboa, Oeiras, Portugal

Viral fusion proteins are essential to allow enveloped viruses (such as Influenza, Dengue, HIV and SARS-CoV-2) to enter their hosts' cells, in a mechanism referred to as membrane fusion. This makes these proteins (with special relevance to their fusion peptides, the component of the protein that can insert into the host's membrane by itself) interesting potential therapeutic targets for preventing or treating for some well-known diseases. However, there is no centralized data repository containing all the relevant information regarding viral fusion proteins.

With that in mind, the main purpose of this work is to develop a CRUD (Create, Read, Update and Delete) web server that will allow researchers to find all the necessary data regarding enveloped viruses and their viral fusion proteins (this data was gathered from biological repositories like NCBI Protein and NCBI Taxonomy, UniProt and PDB), through an easy-to-use web interface. The web application will also contain other bioinformatics functionalities, such as sequence alignment (through BLAST, Clustal and Weblogo) to allow researchers to retrieve key pieces of information regarding a fusion protein, as well as machine learning models capable of predicting the location of fusion peptides inside the viral fusion protein sequence.

The implementation of the server used Django as its back-end, retrieving the data from a MySQL database, and Angular as its front-end.

The main result of the work is, therefore, a working webserver, with a web interface available online through the URL https://viralfp.bio.di.uminho.pt/.

The web application allows users to explore the gathered data related to viral fusion proteins in a user-friendly way. This tool contains all the proposed functionalities and machine learning models. As expected in an application's development, there are several aspects that require future work to improve the usefulness of this tool to the scientific community.

Session 1
6th May
17:00 h

Molecular determinants of the SARS-CoV-2 fusion peptide activity

Carolina C. Buga[1,2], Mariana Valério[1], Alexandra Balola[1], Ana S. Ferreira[1], A M. Sequeira[3], M. Rocha[3], João B. Vicente[1], I. Rocha[1], Miguel A. R. B. Castanho[2], Cláudio M. Soares[1], Ana S. Veiga[2], Diana Lousa[1]

[1] ITQB NOVA, Instituto de Tecnologia Química e Biológica António Xavier, Universidade Nova de Lisboa, Oeiras, Portugal
[2] Instituto de Medicina Molecular João Lobo Antunes, Faculdade de Medicina, Universidade de Lisboa, Lisboa, Portugal
[3] CEB-Centre Biological Engineering, University of Minho, Braga, Portugal

The COVID-19 pandemic, caused by the SARS-CoV-2 virus, emerged in late 2019 and quickly spread worldwide, resulting in over 125 million infections and 2.7 million deaths as of March 2021 accordingly to the World Health Organization. Despite the great advances achieved by the scientific community in providing crucial information about this virus, we are still far from completely understanding it.

SARS-CoV-2 is an enveloped virus, meaning that it is encapsulated by a lipid membrane, which needs to be fused to the host membrane to begin the infection process. Fusion between viral and host membrane is catalyzed by the spike (S) glycoprotein. The S-protein is composed of essential elements for the infection mechanism, namely the receptor-binding domain known to bind to angiotensin-converting enzyme 2 during the viral entry pathway. Another important region, known as the fusion peptide (FP), plays an essential part in the fusion mechanism, by inserting into and disturbing the host membrane. There is still not a consensus among scientists in terms of the fusion peptide location on the S-protein sequence, with two major candidate regions having been proposed.

We recently used a machine learning-based tool developed by us to identify viral FPs with accuracies over 85%. With this tool a putative FP, previously suggested in the literature, has been identified, as well as other proposals including the requirement of more than one FP. To further address this question, we are performing a systematic analysis of the SARS-CoV-2 putative FPs, using Molecular Dynamics (MD) simulations, which provide a detailed perspective of how these peptides insert and interact with the membrane. In parallel, we are characterizing these systems experimentally. Additionally we are exploring therapeutic strategies targeting these regions. Given the major role of the FP in the virus infection process, this work provides relevant insights and contributes to the fight against COVID-19.

Session 1
6th May
17:00 h

PhagePro - prophage finding tool

João Pedro Porto Dias[1], Luís Melo[1], Oscar Dias[1]

[1] Centro de Engenharia Biológica, Escola de Engenharia, Universidade do Minho

Keywords: Bacteriophages, Machine learning, Galaxy, PhagePro

Bacteriophages are viruses that infect bacteria with a high host specificity. As obligate parasites, bacteriophages use the host to reproduce, either by hijacking hosts replication machinery, in a lytic cycle, or by integrating the genetic material in the host's genome, in the lysogenic cycle. The integrated bacteriophage is called prophage and can remain dormant until activated by certain stimuli. These bacteriophage insertions can lead the bacteria to lose or gain functions or have different dynamics with the environment. Prophages can increase bacterial virulence; hence there is the need to study bacteria continuously to understand how the biomes can change and what problems can arise.

Therefore, PhagePro was created to find and classify these alterations. This tool reduces countless *in vivo* procedures, that would be required to isolate and extract bacteriophages. PhagePro can is in four key sections. The first uses machine-learning with additional tools to find putative prophages in the input sequence. The second finds the most probable prophage boundaries, whereas the third searches for protein similarities in Pfam, SwissProt and a bacteriophage protein database. The last section scores the phage by assessing if the phage has a complete life cycle. The tool will be available in Galaxy (https://galaxy.bio.di.uminho.pt/).

During testing, PhagePro demonstrated an excellent performance differentiating bacteriophage sequences from bacterial sequences. Furthermore, when bacterial genomes were tested independently, the algorithm has shown a higher sensitivity than other software predicting putative prophage regions. Further testing will be performed to tune the decision parameters in prophage characterization and scoring.

These results have shown that the PhagePro has great potential in tracking and classifying bacterial evolution mediated by phages. Furthermore, it provides insights on potential alterations that the genome of the bacteriophage may have endured.

Session 2
7th May
10:15 h

Prediction of Gene-Disease Associations through Knowledge Graph Mining

Susana Nunes, Rita T. Sousa, Catia Pesquita

LASIGE, Faculdade de Ciências, Universidade de Lisboa
clpesquita@ciencias.ulisboa.pt

There are still more than 1,400 Mendelian conditions whose molecular cause is unknown. In addition, almost all medical conditions are somehow influenced by human genetic variation. This challenge also presents itself as an opportunity to understand the mechanisms of diseases, thus allowing the design and development of better mitigation strategies, finding diagnostic markers and therapeutic targets. Deciphering the link between genes and diseases is one of the most demanding tasks in biomedical research.

Computational approaches for the prediction of gene-disease associations can greatly accelerate this process, and recent developments that explore the scientific knowledge described in ontologies have achieved good results. State-of-the-art approaches that take advantage of ontologies or knowledge graphs for predicting gene-disease associations are typically based on semantic similarity measures that only take into consideration hierarchical relations. Recent developments in the area of Knowledge Graphs (KG) embeddings support more powerful representations but are usually limited to a single ontology, which may be insufficient in multi-domain applications such as the prediction of gene-disease associations.

We developed a novel approach of gene-disease associations prediction by exploring both the Human Phenotype Ontology and the Gene Ontology, using KG embeddings to represent gene and disease features in a shared semantic space that covers both gene function and phenotypes. Our approach integrates different methods for building the shared semantic space, as well as multiple KG embeddings algorithms and machine learning methods.

The prediction performance was evaluated on curated gene-disease associations from Disgenet and compared to classical semantic similarity measures. Preliminary results show an improvement over classical semantic similarity measures by up to 5% and motivate future work in adapting embeddings strategies to work over multiple linked ontologies.

This work was supported by FCT through LASIGE Research Unit (ref. UIDB/00408/2020 and ref. UIDP/00408/2020) and the project SMILAX (ref. PTDC/EEI-ESS/4633/2014). It was also partially supported by the KATY project which has received funding from the European Union’s Horizon 2020 research and innovation program under grant agreement No 101017453.

Session 2
7th May
10:15 h

Rationally designed antiviral proteins targeting SARS-CoV-2

Susana Parreiras[1], Carlos H. Cruz[1], Mariana Valério[1], Cláudio M. Soares[1], João B. Vicente[1] and Diana Lousa[1]

[1] ITQB NOVA, Instituto de Tecnologia Química e Biológica António Xavier, Universidade Nova de Lisboa, Oeiras, Portugal

SARS-CoV-2 is an enveloped, positive-sense RNA virus that belongs the Coronaviridae family, being responsible for the current global pandemic, COVID-19. So far, no effective treatment against this virus has been discovered.

One of the most promising therapeutic targets of coronaviruses is the spike (S) protein as it is crucial for viral entry, promoting fusion between the viral and host membranes, which allows the virus to insert its genetic material into the host cell and replicate. Although 3D structures of the S-protein of different Coronaviruses are available, the full potential of this protein as a therapeutic target remains poorly explored.

Hence, the aim of this work is to design and produce antiviral proteins that can block the interaction between the S protein and the host receptor, ACE2, and thus prevent SARS-CoV-2 infection.

In a first step several antiviral proteins are computationally designed with the Rosetta program, based on the interactions between ACE2 and the receptor binding domain (RBD) of the S-protein. In the next step, molecular dynamics (MD) simulations of these antiviral proteins (free in solution and in complex with the RBD) are performed to test their structural stability and analyse their interaction with the RBD. Lastly, experimental validation is carried out in order to ascertain their structural and conformational stability, antiviral properties, and binding affinity for the S protein. The current results show that the designed proteins are stable and form extensive contacts with the RBD, which indicates that these proteins are promising therapeutic candidates to fight COVID-19.

Session 1
6th May
17:00 h

Stoichiometric genome-scale models for the chondroitin production in *Escherichia coli*

Márcia R. Couto[1], Oscar Dias[1], Joana L. Rodrigues[1], Lígia R. Rodrigues[1]

[1] Centre of Biological Engineering, University of Minho, Campus de Gualtar, 4710-057 Braga, Portugal

Keywords: Chondroitin, *Escherichia coli* BL21(DE3), *in silico*, metabolic engineering, stoichiometric genome-scale model, heterologous pathway.

Chondroitin is a natural-occurring glycosaminoglycan with applications as a nutraceutical and pharmaceutical ingredient. It can be extracted from animal tissues, though chondroitin-like polysaccharides using microorganisms emerged as a safer and more sustainable alternative source. However, chondroitin yields using either natural or recombinant microorganisms are still far from meeting the increasing demand. In this work, stoichiometric models containing the heterologous pathway necessary for producing chondroitin in *E. coli* were constructed and investigated for mutant predictions that would potentially improve chondroitin yields. Four models of *E. coli* BL21 (BIGG ID: iECBD_1354, iECD_1391, iEC1356_Bl21DE3, iB21_1397) and one of *E. coli* K12 (BIGG ID: iJO1366), from which the other models were derived, were used to insert the heterologous pathway composed by two enzymatic steps catalyzed by UDP-*N*-acetylglucosamine 4-epimerase (UAE) and chondroitin synthase/polymerase (CHSY). The models were imported in Optflux, and the evolutionary optimization was then performed for gene deletion predictions using Strength Pareto Evolutionary Algorithm 2 (SPEA2) and the parsimonious Flux Balance Analysis (pFBA) as the simulation method. Chondroitin production was not predicted to improve by combining gene deletions, probably because the competing pathways that use the intermediates are critical for cell growth. However, gene over and underexpression search allowed to identify several targets. Most of the resulting solutions were composed by the overexpression of one of the genes responsible for the production of the heterologous pathway precursor (either *glmU* or *glmM* encoding glucosamine-1-phosphate *N*-acetyltransferase/UDP-*N*-acetylglucosamine diphosphorylase and phosphoglucosamine mutase, respectively) combined with the underexpression of one of the genes associated with cell wall recycling pathways (such as membrane-bound lytic transglycosylases *mltA*, *mltB* and *mltC*, or the anhydromuropeptide permease *ampG*), which contain reactions known to consume such precursors. The solutions herein obtained will be further validated *in vivo* by constructing the *E. coli* mutants predicted to improve chondroitin production.

Márcia R. Couto acknowledges her grant (SFRH/BD/132998/2017) from the Portuguese Foundation for Science and Technology (FCT). This study was supported by the FCT under the scope of the strategic funding of UIDB/04469/2020 unit.

Session 2
7th May
10:15 h

The BioData.pt Microbiome Portal: A Platform to Unify Portuguese Microbiota Research

João Rato[1], Ricardo Leite[1], Rafael Santos[1], Ana Melo[1], and Daniel Faria[2]

[1] Biodata.pt, Instituto Gulbenkian de Ciência (IGC), Oeiras, Portugal

[2] BioData.pt, Instituto de Engenharia de Sistemas e Computadores - Investigação e Desenvolvimento (INESC-ID), Lisboa, Portugal

Microbiome research has grown considerably during the past few decades and that has resulted in the generation of huge amounts of data raising challenges regarding its alignment with the FAIR principles (Findability, Accessibility, Interoperability and Reusability). BioData.pt, the Portuguese Node of ELIXIR, assembled a national community to address microbiome research questions, particularly in what refers to data analysis and management, an effort that is also being chased by our European counterpart.

To unify and fairify all data generated by microbiota research in Portugal we are developing the BioData.pt Microbiome Portal. This portal is built on the PERN stack which consists of PostgreSQL, Express, React and Node.js. The approach used for the database construction uses a sparse matrix foreseeing required and open metadata to allow the best description of the data.

The BioData.pt Microbiome Portal will support researchers in the management of their data by providing metadata standards for data deposition and shareability with the creation of working groups with different levels of access and visibility for each project. Furthermore, it will also give the user the option to submit their data to international repositories and make accessory tools available for metagenomics data analysis. The latter will enable the user to choose between different pipelines and databases, including custom databases, which will be deposited in the platform once the analysis is completed.

Session 2
7th May
10:15 h

The natural history of clonal haematopoiesis

Margarete Fabre[1,2*], José Guilherme de Almeida[3*], Edoardo Fiorillo[4], Valeria Orru[4], MS Vijayabaskar[2], Joanna Baxter[5], Claire Hardy[1], Federico Abascal[1], Iñigo Martincorena[1], Eoin McKinney[5], Francesco Cucca[6], Moritz Gerstung[3†], George Vassiliou[1,2,7,8†].

[1] Wellcome Sanger Institute, Wellcome Genome Campus, Cambridge, CB10 1SD, UK

[2] Wellcome-MRC Cambridge Stem Cell Institute, University of Cambridge, Cambridge, CB2 0XY, UK

[3] European Bioinformatics Institute, Wellcome Genome Campus, Cambridge, CB10 1SD, UK

[4] Cambridge Blood and Stem Cell Biobank, Department of Haematology, University of Cambridge, Cambridge, CB2 OAW, UK

[5] Cambridge Institute of Therapeutic Immunology & Infectious Disease, University of Cambridge, Cambridge, CB2 OAW, UK

[6] Istituto di Ricerca Genetica e Biomedica, Consiglio Nazionale delle Ricerche, Monserrato, Italy

[7] Cyprus Cancer Research Institute, University Avenue, 2109 Nicosia, Cyprus

[8] Department of Haematology, Cambridge University Hospital NHS Foundation Trust, Cambridge, CB2 0QQ, UK.

[*] These authors contributed equally to the work

Clonal haematopoiesis (CH) describes the clonal expansion of mutant haematopoietic stem cells in healthy and mostly elderly individuals. Yet CH is also considered an important step in the progression to blood cancers, such as acute myeloid leukaemia and myelodysplastic syndrome. Therefore, understanding CH evolution is important to better understand the progression from seemingly inofensive somatic mutation to cancer. In this work, we use longitudinal deep targeted sequencing data for 385 elderly individuals to estimate clonal dynamics and age at onset for all clones. We observe distinct annual growth rates for each gene, with the lowest occurring in DNMT3A (4%) and the highest in U2AF1 (21%). The extrapolated age at onset of clonal growth shows similar patterns for most genes, with clones arising uniformly through life. Against this trend, mutations such as those in U2AF1 or SRSF2-P95H only appear much later in life. With single cell colonies for 3 individuals we are able to verify our estimates and observe clonal expansions with no known drivers, suggesting that a portion of CH goes by unnoticed. This work offers a comprehensive view on the dynamics and evolution of CH, disentangling the effects that driver genetics and other effects have on CH progression.

Session 1
6th May
17:00 h

The Structural and Functional Role of a CACNG2 Mutation in Psychiatric Disorders

Raquel P. Gouveia[#a], Carlos A. V. Barreto[#a,b], Salete J. Baptista[a,c], Gladys Caldeira[a,b], A. J. Preto[a,b], Rita Melo[a,c], Ana L. Carvalho[*a,d], and Irina S. Moreira[*d,a]

[a] Center for Neuroscience and Cell Biology. University of Coimbra, UC Biotech Building, 3060-197 Cantanhede, Portugal.
[b] Institute for Interdisciplinary Research, University of Coimbra, Coimbra, Portugal
[c] Universidade de Lisboa, Estrada Nacional 10, ao km 139,7; 2695-066 Bobadela, Portugal.
[d] Coimbra University, Department of Life Sciences, Center for Neuroscience and Cell Biology, 3000-456 Coimbra, Portugal.
[#] co-first authors
[*] corresponding authors

Neuropsychiatric disorders incidence has increased worldwide and became an increasing economic and social burden. Intellectual disability and schizophrenia, key examples of such disorders, were already associated with proteins related to homeostatic plasticity. Even though the number of proteins associated with these conditions is massive, the mechanism that leads to the disease is still poorly understood.

Interestingly, various studies showed that the same gene may be found mutated in patients diagnosed with different neuropsychiatric disorders. This may point to a common origin related to synaptic dysfunction.

The CACNG2 gene, which codes for stargazing (STG), was described as a susceptible gene for psychiatric disorders. STG, an auxiliary subunit for α-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid receptors (AMPAR), plays an important role in modulating AMPAR function, transporting it to the synapse and assisting in the homeostatic synaptic scaling of AMPAR. There is already one key mutation in CACNG2 reported in literature linked to intellectual disability.

To better understand the effects of this mutation on the structure of STG and in its interaction with AMPAR, in silico modeling techniques were used to obtain models of the structures of both the STG (with and without the mutation) and AMPAR, followed by molecular dynamics simulations. Structural and dynamical analysis, such as ΔSASA, MMPBSA calculations and interface interactions (H-bonds and salt-bridges), throughout the simulation time unravel differences in the interface between STG and AMPAR. This atomic-level information might explain the reported hindered function of mutant STG.

Session 1
6th May
17:00 h

Understanding the genome architecture and evolution of Shiga toxin encoding bacteriophages of *Escherichia coli*

Graça Pinto[1,2], Marta Sampaio[1], Óscar Dias[1], Carina Almeida[2], Joana Azeredo[1], Hugo Oliveira[1]

[1] CEB - Centre of Biological Engineering, University of Minho, 4710-057, Braga, Portugal
[2] INIAV, IP-National Institute for Agrarian and Veterinary Research, Rua dos Lagidos, Lugar da Madalena, Vairão, Vila do Conde, Portugal

Shiga toxin-producing *Escherichia coli* (STEC) is an important foodborne pathogen, and its major virulence factor is their ability to produce Shiga toxins. This toxin is coded by the *stx* gene, acquired through the insertion of a prophage into their genome. In our study, 179 STEC genomes were analysed for their serotype, distribution, and *stx* gene variants. Stx phages were also analysed and grouped based on shared gene content. We show that most STEC were isolated from different sources and geographical regions and belong to the non-O157 serotypes (73%). While the majority of STEC encode a single *stx* gene (61%), strains coding for two (35%), three (3%) and four (1%) *stx* genes were also found, being *stx2a* the most prevalent gene variant. PHASTER analysis found *stx* genes in intact prophage regions, indicating they are phage-borne. Stx phages from our dataset were grouped into four clusters (A, B, C and D), three subclusters (A1, A2 and A3) and one singleton, in agreement with the predicted virion morphologies. Stx phage genomes are highly diverse with a vast number of 1,838 gene phamilies (phams) of related sequences (of which 677 are orphams i.e. unique genes) and although having high mosaicism, they are generally organized into three major transcripts (structural, metabolism, lysis and virulence). There is a strong selective pressure to maintain the *stx* genes location in close proximity to the lytic cassette composed of predicted SAR-endolysin and pin-holin lytic proteins. Taken together, we demonstrate that Stx phages' genomes are highly diverse, with several lysis-lysogeny regulatory systems identified but with a conserved lytic system always adjacent to *stx* genes.

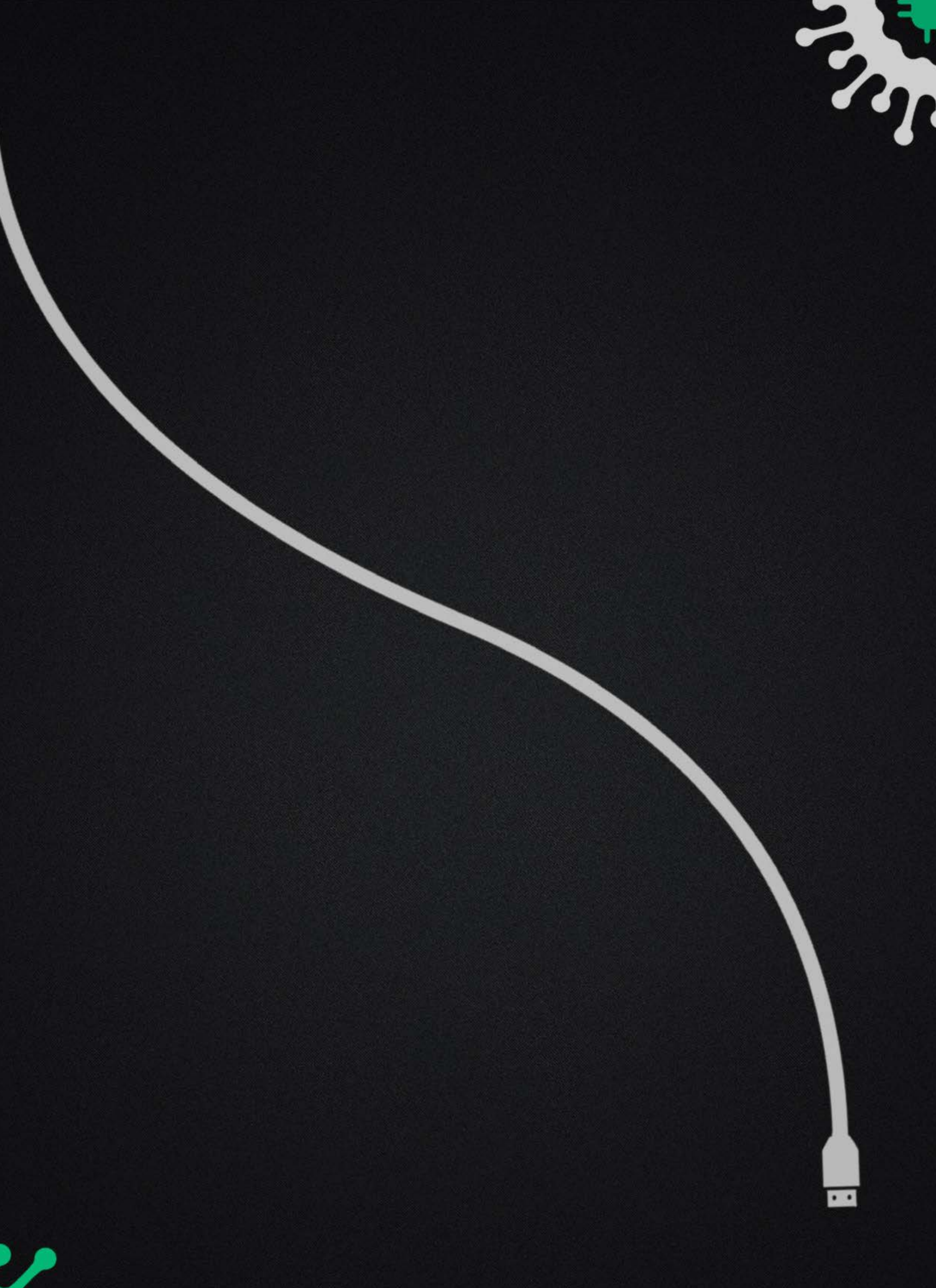
Session 1
6th May
17:00 h

UPIMAPI, reCOGnizer and KEGGCharter: three tools for functional annotation

João C. Sequeira[1], Miguel Rocha[1], M. Madalena Alves[1], Andreia F. Salvador[1]

[1] CEB-Centre of Biological Engineering, University of Minho, Braga, Portugal

Omics technologies generate large datasets from which biological information must be extracted by using bioinformatics tools. Although web services provide easier to use interfaces, large datasets are difficult to handle. This is not a limitation of command-line tools and programmatic modules, but these may be challenging to use. In this work, three command-line tools were developed, aimed for speed and automation. The tools are available through Bioconda for Unix systems and were developed in Python 3, making use of multithreading/multiprocessing in computationally demanding steps. UPIMAPI integrates annotation with reference to the UniProt database with automatic retrieval of internal and cross-reference information from other databases (e.g., KEGG, BRENDA and RefSeq) through UniProt's API, accessed with urllib package. The input is a FASTA file containing protein sequences, and the outputs are EXCEL or TSV files containing taxonomic, functional, and cross-reference information. reCOGnizer performs domain-based annotation of protein sequences with CDD, Pfam, NCBIfam, Protein Clusters, TIGRFAM, SMART, COG and KOG as reference databases, and obtains EC numbers and taxonomic assignments per domain identified. The results are outputted in TSV and EXCEL files. KEGGCharter is a command line implementation of KEGG Pathway's mapping service, while also obtaining additional KOs and EC numbers, through the methods available in BioPython for accessing KEGG's API. KEGGCharter takes as input a table (TSV or EXCEL), containing either KEGG IDs, KOs or EC numbers. KEGGCharter represents identified KOs in metabolic maps and includes information on differential gene expression. When data from more than one organism is uploaded, KEGGCharter links function to taxonomic identification, which can be visualized in the maps. Differential expression of genes/proteins can be visualized in metabolic maps, by showing mini heatmaps. UPIMAPI and reCOGnizer are complementary tools, providing functional annotation based on protein sequencing homology and on identification of protein conserved domains, respectively. Both tools retrieve the IDs (KEGG IDs, EC numbers and KOs) necessary to run KEGGCharter. Together, these tools provide a complete characterization and visualization of results, facilitating the interpretation of omics experiments, and requiring minimal bioinformatics expertise.



5-8 MAY 2021