XI EDITION

3-5th March

CONFERENCE BOOK

Universidade do Minho
Escola de Engenharia

WELCOMING_MESSAGE

Greetings,

On behalf of the Organisers of this event, it is a great pleasure to welcome you to the 11th edition of Bioinformatics Open Days. It is a great privilege to welcome this diversity of students, researchers and academics who participate in our event and come to showcase some of the exceptional work that has been produced in the field of Bioinformatics.

Thanks to the significant growth within Bioinformatics, both in the scientific and technological fields, enormous challenges and opportunities have arisen both in the business and scientific fields. There has also been remarkable growth in Portugal over the last few years. Thanks to this growth, prestigious post-graduate courses have been developed at the academic level, and at the business level, new start-up companies with international connections have emerged.

Bioinformatics Open Days is an initiative held since 2012 at the University of Minho in Braga. Promoted and led by students, this initiative aims to promote the exchange of knowledge between students, professors, and researchers in the areas of Bioinformatics and Computational Biology, describing the present and future of Bioinformatics, both nationally and internationally.

Fortunately, this year we had the opportunity to make the event in the face-to-face format again, keeping our commitment to provide an event with the quality that has been offered in recent years.

Finally, we would like to welcome all the participants of this event, hoping that you can acquire and share valuable knowledge.

The organisation of BOD 2022

THE_ORGANIZING_COMMITTEE

Miguel Rocha
General Chair

Maria Couto

João Monteiro

Tiago Machado

Mariana Coelho

Diogo Macedo

Joana Gabriel

Sara Boaventura

Teresa Coimbra

Tiago Silva

Rui Gomes

André Magalhães

Diogo Cachetas

Mariana Pereira

Miguel Barros

Alexandre Castro

Juliana Pereira

Mónica Fernandes

Alexandre Esperança

Carla Silva

José Pereira

Luís Moncaixa

TABLE_OF_CONTENTS

PROGRAM

Thursday, 3rd March

09:00 Check-in
Anfiteatro CP2 B1

09:30 Opening Session
Anfiteatro CP2 B1

10:00 Keynote - Quantifying gut microbiota-drug-host metabolic interactions
Dr. Maria Zimmerman
Anfiteatro CP2 B1

10:45 Coffee Break
CP2

11:15 Oral Communications - Health Applications
Anfiteatro CP2 B1

OC1 - Exploring new algorithms to combine genomic information from extant populations and ancient specimens | Joana C. Ferreira, Farida Alshamali, Luisa Pereira, Veronica Fernandes

OC2 - A Validated Proteomics and Genomics Pipeline to Unravel Prostate Cancer Biomarkers | Tânia Lima, Rita Ferreira, Marina Freitas, Rui Henrique, Rui Vitorino, Margarida Fardilha

OC3 - Whole genome sequencing analysis: exploring germline CNV and SNV landscapes | Marta Ferreira, Joaquin Jurado Maqueda, Francisco Almeida, Rita Monteiro, Celina São José, Carla Oliveira

OC4 - Understanding HIV-1 epidemiology and genetic diversity in Brazil | Ana Santos Pereira, Triunfante V.; Souto B.; Araújo P. M. M., Martins J.; Osório N. S

13:00 Lunch Break

14:30 Keynote - Common therapeutics induce emergent behaviours in synthetic gut bacterial communities
Dr. Sarela Garcia-Santamarina
Anfiteatro CP2 B2

15:15 Oral Communications - Machine Learning
Anfiteatro CP2 B1

OC5 - Stable variable selection in penalised regression models: an application to high dimensional genomic data - Alzheimer's Disease | Leonor Rodrigues, Ana H. Tavares, Vera Enes, Miguel Pinheiro, Gabriela Moura, Vera Afreixo

OC6 - Knowledge Graph Embeddings for ICU readmission prediction | Ricardo Carvalho, Daniela Oliveira, Catia Pesquita

OC7 - Transcriptomic effects of cigarette smoking across human tissues | Rogério Ribeiro, Raquel García-Pérez, José Cardenosa, Marta Melé & Ferreira P. G.

16:15 Poster Presentations / Coffee Break
CP2

16:45	Oral Communications - Bioinformatics Applications *Anfiteatro CP2 B1* OC8 - New Insights into the Catalytic Mechanism of Plastic Degrading Enzyme IsPETase: a QM/MM computational study | Rita P. Magalhães, Henrique S. Fernandes, and Sérgio F. Sousa OC9 - The catalytic mechanism of Pdx2 glutaminase: a QM/MM approach | André F. Pina, Sérgio F. Sousa, Nuno M. F. S. A. Cerqueira OC10 - Bioinformatic strategies in grapevine genomics, from biodiversity to domestication | Sara Freitas, Miguel Carneiro, Herlander Azevedo
17:45	End of the Day
19:00	Speed Meeting
20:45	Formal Dinner *Migaitas Restaurant*

Friday, 4th March

10:00	Keynote - Streamlining the Design-Build-Test-Learn workflow for bio-based production platforms Dr. Pablo Carbonell *Anfiteatro CP2 B1*
10:45	Poster Presentation / Coffee Break *CP2*
11:15	Oral Communications – Systems Biology *Anfiteatro CP2 B1* OC11 - Drug Target Identification based on the construction of a Genome-scale metabolic model for the human pathogen Candida parapsilosis | Diogo Couceiro, Romeu Viana, Tiago Carreiro, Oscar Dias, Isabel Rocha, Miguel Cacho Teixeira OC12 - Strain optimisation for aromatic amino acids using an Escherichia coli kinetic model | André Fonseca, Isabel Rocha OC13 - A continous approach for modeling fermentation phases in wine production | Artai Rodríguez-Moimenta, David Henriques, Eva Balsa-Canto OC14 - MEWpy: Modelled by Man, designed by Nature | Vítor Pereira
13:00	Lunch Break
14:30	Companies Presentations *Anfiteatro CP2 B1*
15:00	Networking Sessions *CP2*
16:00	Round Table *Anfiteatro CP2 B1*
16:45	Award Ceremony
17:30	Social Activities

19:00	Informal Dinner *Bar Carpe*

Saturday, 5th March

09:30	Workshops Workshop 1: Machine and deep learning applied to protein sequences | Ana Marta Sequeira Workshop 2: Single-cell RNA Sequencing - One cell at a time | Ana Falcão, Mónica Fernandes, Diogo Macedo
12:00	End of the Day

SPEAKERS

DR. MARIA ZIMMERMANN-KOGADEEVA

Dr. Maria Zimmermann-Kogadeeva is the leader of the Multi-omics-based modelling of microbial ecosystems group at EMBL Heidelberg. After completing her PhD at ETH Zurich, with emphasis on machine learning-based methods for generalised and high-throughput 13C metabolic flux analysis to study bacterial adaptations to complex environments, Dr. Zimmermann-Kogadeeva worked as a postdoctoral associate at Yale University School of Medicine. Currently, as stated before, Dr. Zimmermann-Kogadeeva is one of the newest group leaders at EMBL Heidelberg focusing her research on the metabolic interactions between gut microbes, their environment, and their host. Her group is currently developing mathematical models capable of describing the underlying processes based on time-resolved and spatial metabolomic measurements.

DR. SARELA GARCÍA-SANTAMARINA

After completing her PhD in Biomedicine at University Pompeu Fabra, Spain in 2013, Dr. Garcia-Santamarina is now an Auxiliary Investigator at MOSTMICRO- ITQB Nova as group leader. Her interests shift towards the importance of the human microbiome for health and disease with the end goal of contributing to a better understanding of microbial physiology and responses to fluctuating micronutrient concentrations, as well as shed light on host-microbial interactions that might have a direct role in host nutrition. During her research career from EMBL-Heidelberg to Universidade de Santiago de Compostela, she counts with numerous relevant published works.

DR. PABLO CARBONELL

Dr. Pablo Carbonell is Senior Reader in Computational Biology at Universitat Politècnica de València, holding a BSc and MSc in Industrial Electronics Engineering, a MSc in Bioinformatics, and a PhD in Control Engineering and Research Habilitation in Systems Biology. Dr. Carbonell conducted research in numerous prestigious research centers. Currently, his research is focused on automated design for metabolic engineering and synthetic biology. He has contributed to the development of several bioretrosynthesis-based pathway design tools and theoretical models for bio-based systems.

ROUNDTABLE

In the digital Revolution Era we live in, value creation from technology is paramount. Hence, as present and future workers in the biological, digital, and technological fields, it is essential to learn more about the industry's growth trends and future challenges.

In this networking session, we will be joined by four cutting-edge leading companies: Accenture, Amyris, CBR Genomics and iLoF. These companies work in various biological and technological areas ranging from health applications, personalised medicine and systems biology to digital transformation, consulting, and data science.

In these ever-developing areas, it is crucial to adapt to the operational challenges and have collaborative thinking to overcome such challenges.

Join us at this session to debate the paths in these industries, their goals, future obstacles, challenges and what we can do to address them.

accenture

amyris

cbrgenomics
genetically informed medicine

iLoF

NETWORK_SESSION

ACCENTURE

Accenture is a global professional services company with leading digital, cloud, and security capabilities.

Combining unmatched experience and specialised skills across more than 40 industries, the company offers strategy and consulting, interactive, digital, technology and operations services—all powered by the world's largest network of Advanced Technology and Intelligent Operations centers. Accenture's people deliver on the promise of technology and human ingenuity every day, serving clients in more than 120 countries.

AMYRIS

Amyris aims to deliver the lowest cost, most disruptive way to produce pure ingredients from sustainable sources with no compromise. The company was founded in 2003 with a vision to remake the world's chemistry using synthetic biology by genetically engineering yeast and fermenting it with sustainably-sourced sugarcane to produce natural, high-performance molecules. Today, they have manufactured and sold more bio-based sustainable products from fermentation than all companies in their sector. This company's goal is to shift the world into sustainable ingredients that will lead to other global shifts.

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 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 there is a daily work 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.

iLOF

iLoF is empowering a new era of personalised 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.

WORKSHOPS

Machine and Deep Learning applied to protein sequences

Ana Marta Sequeira

In this workshop, introductory topics related to the application of machine and deep learning to protein sequences will be explored using Python. Different approaches to encode an aminoacid sequence as a numerical vector will be explored. The workshop will also focus on analysing datasets of proteins resorting to unsupervised exploration and Machine and Deep Learning techniques.

Single-cell RNA Sequencing – One cell at a time

Ana Falcão

Mónica Fernandes

Diogo Macedo

In this workshop we will introduce scRNA-seq technique and explore the current methods for processing and analysing scRNA-seq datasets. The pipeline that will be here presented starts from the read count matrices to cell trajectories and other advanced insights.

Health applications

Exploring new algorithms to combine genomic information from extant populations and ancient specimens

Joana C. Ferreira[1,2,3], Farida Alshamali[4], Luisa Pereira[1,2], Veronica Fernandes[1,2]

[1]i3S—Instituto de Investigação e Inovação em Saúde, Universidade do Porto, Portugal;
[2]IPATIMUP—Instituto de Patologia e Imunologia Molecular da Universidade do Porto, Portugal;
[3]ICBAS—Instituto de Ciências Biomédicas Abel Salazar, Universidade do Porto, Portugal; 4Dubai Police General Headquarters, United Arab Emirates

Coalescence analyses on modern-day genomes have been fundamental for inferring major events of the human past, but the recent boom in ancient DNA (aDNA) is revealing a more complex picture, with many coexisting and interbreeding groups, some surviving and other becoming extinct. This aDNA information can be projected in the modern genomes, allowing the statistical evaluation between alternative hypotheses for their origins. In order to be able to do so, it was necessary to develop new algorithms that combine genomic information from extant populations and ancient specimens.

We will illustrate the power of this approach by combining our dataset of 741,000 variants screened in modern-day 291 Arabians and 78 Iranians with 304 available aDNA (including Levant and Caucasus/Iran specimens as old as 15,000 years). Our aim was to contribute insights into the role played by Arabian Peninsula as one of the first places receiving the out-of-African (OOA) migrants, at ~60,000 years ago, in the path for the settlement of the entire globe. We first projected aDNA diversity in the principle component (PCA) and ADMIXTURE profiles of modern genomes, to identify the best potential founders for Arabians. Then, we modelled specific ancestry scenarios by using qpAdm, CP/NNLS and f4-statistic algorithms. Our analyses detected a signature of ancient ancestry (known as basal Eurasian) in the eastern part of the Peninsula, in the Arabo-Persian Gulf, that by the OOA time was an exposed basin. Even today, eastern Arabs show the highest levels of this ancestry amongst extant populations. These basal Eurasians were genetically close to another group from Levant, but while those did not interbreed with the Neanderthals, these had limited (or restricted in time) interbreeding with this archaic group. Interestingly, it was the group that mixed with Neanderthals that was in the origin of all current Europeans and Asians, whom inherited a mean of 3% Neanderthal input in their genomes.

Funding:

This work was financed by FEDER funds through the COMPETE 2020 - Portugal 2020 (POCI-01-0145-FEDER-016609) and by Portuguese funds through FCT - (P2020- PTDC/IVC-ANT/2421/2014) in the framework of the project "Biomedical anthropological study in Arabian Peninsula based on high throughput genomics“. The author V. Fernandes has a post-doc grant through FCT (SFRH/BPD/114927/2016).

A Validated Proteomics and Genomics Pipeline to Unravel Prostate Cancer Biomarkers

Tânia Lima[1,5], Rita Ferreira[2], Marina Freitas[1], Rui Henrique[4,5,6], Rui Vitorino[1,2,3], Margarida Fardilha[1]

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

Prostate cancer (PCa) is the most prevalent noncutaneous cancer among men, but when detected in early stages it can save many lives. However, due to the limited accuracy and invasive nature of current diagnostic tools, diagnosing PCa is challenging. Thus, the discovery of robust and noninvasive biomarkers for PCa is of paramount importance. In that vein, an innovative approach was taken in the PCa context based on an automatic text-mining feature of VOSviewer and integrated proteomic and genomic analysis. VOSviewer was used to retrieve and create co-occurrence networks of terms associated with PCa. These results were complemented with DisGENET data, and with a recent bioinformatic analysis integrating all differentially expressed proteins identified in tumor tissue and urine from PCa patients. Afterward, the results were integrated with gene expression data from the Gene Expression Omnibus database to correlate gene and protein levels. This study suggests AXIN2, GSTM2, KLK3, LGALS3, MSMB, PRTFDC1, and SH3RF1 as important entities in PCa context. KLK, LGALS3, and MSMB proteins are common to a previous bioinformatic analysis, and a concordance was found between their gene and protein expression levels. With the aim of validating the applicability and reliability of the pipeline presented in this manuscript, the galectin-3 (LGALS3) protein, whose function in PCa is underexplored, was selected for validation. Urinary levels of galectin-3 were assessed in samples from PCa patients and non-cancer subjects and were shown to vary between groups. This finding helps to better understand PCa biology and confirm the applicability of this pipeline in the discovery of new biomarkers.

Whole genome sequencing analysis: exploring germline CNV and SNV landscapes

Marta Ferreira[1,2], Joaquin Jurado Maqueda[1,2], Francisco Almeida[1,2,5], Rita Monteiro[1,2,4], Celina São José[1,2], Carla Oliveira[1,2,3]

[1]Expression Regulation in Cancer, Institute of Molecular Pathology and Immunology of the University of Porto (IPATIMUP), 4200-135 Porto, Portugal
[2] Instituto de Investigação e Inovação em Saúde (i3S), University of Porto, 4200-135 Porto, Portugal
[3]Dept. Pathology and Oncology, Faculty of Medicine, University of Porto, 4200 - 319 Porto, Portugal
[4]Currently at: Inovretail, SA, 4200-355 Porto
[5]Dept. Biology, Faculty of Science, University of Porto, 4169-007 Porto, Portugal

Genomes are naturally prone to aberrations, such as single nucleotide variants (SNVs) or structural variants (SV), which need continuous correction. Those that remain uncorrected, likely explain part of the missing heritability that exome analysis could not reveal, particularly in non-coding regions of disease-causing genes. Genome-wide analysis is therefore pivotal to understand genome variation, and Whole Exome Sequencing (WES) and Whole Genome Sequencing (WGS) are the preferred methods for that purpose. Among other considerations, a lack of user-friendly data analysis and interpretation tools are the major barrier to routine use of these techniques. Herein, we present the main advantages of WGS in relation to other approaches, as well as the steps of an analysis pipeline designed to study WES and WGS. We combined a complete set of tools and infrastructures to deliver the interpretation of WES/WGS. Our proposed WES/WGS analysis workflow includes, besides conventional alignment and post-processing, a combination of multiple variant callers for both CNV and SNV, which were submitted to overlap analysis to improve the pick-up rate of true positives.

The work that we have been developing allowed us to understand the main problems in analysing WES/WGS and how to solve them. With this pipeline, we will be able to analyse WES/WGS fast and accurately and give opportunity to integration into other pipelines as well as the possibility to have a readout easy to understand for the health-practitioners.

The authors declare no conflicts of interest.

Funding:

SOLVE-RD (H2020-SC1-2017, funded by Horizon 2020), European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 722148, FCT Fellowships (2020.05763.BD to MF)

Understanding HIV-1 epidemiology and genetic diversity in Brazil

Santos-Pereira A., Triunfante V.; Souto B.; Araújo P. M. M., Martins J.; Osório N. S.

[1] Life and Health Sciences Research Institute (ICVS), School of Medicine, University of Minho, Braga, Portugal
[2] ICVS/3B's - PT Government Associate Laboratory, Braga, Guimarães, Portugal
[3] Department of Medicine, Federal University of São Carlos, São Paulo, Brazil

In 2019, 38 million people were estimated to be living with Human Immunodeficiency Virus (HIV) infection. There is still no cure for HIV infection and the extensive viral genetic diversity represents the one of the biggest obstacles to develop an effective therapy. Brazil represents the country with the largest number of people living with HIV in Latin America, presenting a heterogeneous distribution of HIV-1 subtypes and recombinant forms across different geographic regions. Understanding HIV-1 epidemics in a country as vast as Brazil, could give valuable insights about the viral genetic diversity and its relationship with genetic and sociodemographic characteristics of the host population. Thus, we firstly aimed at understanding the dynamics of the surveillance drug resistance mutations (SDRMs) in HIV-1 infected individuals failing antiretroviral treatment (ART) in Brazil, between the years 2008 and 2017, and, posteriorly, at investigating the molecular epidemiology and evolution of HIV-1 subtypes, in the same country. Our results revealed a high prevalence of SDRMs in the studied population, although a mild decline was observed over the years. A significant increase on the prevalence of the K65R reverse transcriptase mutation was noticed, following a shift on the used ART regimens. Evidence of K65R transmission was also verified and our results suggested that this mutation could enhance viral recognition by HLA-B27 that has relatively low prevalence in the Brazilian population. Moreover, our results indicated an increase on subtype C prevalence over the years, especially in the South of Brazil. We also observed evidence for subtype C transmission events between the South and other Brazilian regions, although this subtype was only present in small proportions in these areas. Additionally, subtype C was significantly associated with lower levels of immunodepression infection of women and women-to-child transmission, when compared with subtype B, sustaining the hypothesis that some subtypes might take advantage of longer asymptomatic periods and the sociodemographic characteristics of the population to proliferate. Overall, our results reinforce the importance of understanding the dynamics of HIV-1 subtype expansion and monitoring SDRMs prevalence to establish specific guidelines for prevention and treatment, aiming at decreasing the epidemic burden of the HIV-1 infection.

Machine Learning

Stable variable selection in penalised regression models: an application to high dimensional genomic data - Alzheimer's Disease

Leonor Rodrigues1, 2, Ana H. Tavares2, 3, Vera Enes1 , Miguel Pinheiro4, 5 , Gabriela Moura4, 5, Vera Afreixo1, 2

1 Department of Mathematics, University of Aveiro
2 Center for Research & Development in Mathematics and Applications - CIDMA
3 Águeda School of Technology and Management - ESTGA
4 Department of Medical Sciences, University of Aveiro
5 Institute of Biomedicine – iBiMED

Introduction: Alzheimer's disease (AD) is a complex disorder caused by a combination of environmental and genetic factors [1]. One of the main goals of modern genetics has been unravelling the genetic background of common complex disorders. The challenge in finding a plausible method to apply in genetic data is due to its high dimensionality. This work aims to find a consistent method that combines penalised regression techniques and Akaike's Information Criterion (AIC) that identifies a correlation between some Single Nucleotide Polymorphisms (SNPs) and AD.

Methods: The data involved in this study was obtained from the Alzheimer's Disease Neuroimaging Initiative – 1 public database. In this work, prediction models were constructed using Least Absolute Shrinkage and Selection Operator ($\alpha = 1$) and Elastic-net ($\alpha \in \{0.75, 0.50, 0.25, 0.10, 0.05, 0.01\}$). Two types of models were considered: one only with SNPs and another with SNPs but adjusted to age, sex, education level and APOE4 covariates. For each type of model, 100 models were constructed for each value of α. Two final models were proposed for each α: each final model only contains the predictor variables considered important (weight based on AIC being at least 0.8). The performance of these models was measured using Area Under the Curve, Accuracy and F1-measure.

Results: It was found a better performance when $\alpha = 1$ in the case of the models considering only the SNPs. Whereas the opposite was noticed in models adjusted to covariates in which the model with $\alpha = 0.01$ demonstrated the best performance. In both models, the selected variables are similar. There is a consistency of the results for the common SNPs.

Conclusions: The rs2075650 SNP was selected in the two best models proposed. The APOE4 gene was selected in the model adjusted to covariates. Both variables were significant and considered risk factors. They are already reported in the literature as risk factors for AD [2], [3].

References

[1] P. G. Ridge, S. Mukherjee, P. K. Crane, and J. S. K. Kauwe, "Alzheimer's disease: Analysing the missing heritability," PLoS One, vol. 8, no. 11, pp. 1–10, 2013, doi:10.1371/journal.pone.0079771.
[2] H. Huang et al., "The TOMM40 gene rs2075650 polymorphism contributes to Alzheimer's disease in Caucasian, and Asian populations," Neurosci. Lett., vol. 628, pp. 142–146, 2016, doi: 10.1016/j.neulet.2016.05.050.
[3] H. Stocker, T. Möllers, L. Perna, and H. Brenner, "The genetic risk of Alzheimer's disease beyond APOE ε4: systematic review of Alzheimer's genetic risk scores," Transl. Psychiatry, vol. 8, no. 166, 2018, doi: 10.1038/s41398-018-0221-8.

Knowledge Graph Embeddings for ICU readmission prediction

Ricardo Carvalho, Daniela Oliveira, Catia Pesquita

LASIGE, Faculdade de Ciências, Universidade de Lisboa racarvalho@fc.ul.pt

Readmissions to intensive care units (ICU) are a critical problem, associated with either serious conditions, illnesses, or complications (Lai et al., 2012). In normal conditions, ICU readmissions are a burden to the patients representing a 4 times increase in mortality risk, and a financial burden to the health institutions because readmission costs about 20% of the full hospital stay. In developed countries 1 in every 10 patients discharged comes back to the ICU but if multiple co-morbidities are factored, this reaches 1 in every 8 patients (Correa et al., 2017).

In recent years new efforts emerged, using machine learning approaches to make readmission predictions directly over ICU patient's data in formats like Electronic Health Records (EHR) data (Huang et al., 2019; Scherpf et al., 2019; Suresh et al., 2018), but sophisticated machine learning approaches achieve relatively limited improvements over simpler ones. Biomedical ontologies and knowledge graphs provide a means to contextualise EHR data with existing scientific knowledge, but existing works fail to explore it.

We integrated knowledge graph embeddings which provide vector representation of the semantics of data within an ICU readmission prediction pipeline and evaluated whether readmission prediction is improved, and how different knowledge sources impact prediction.

The proposed methodology includes the semantic annotations of the EHR data with different ontologies, the generation of embeddings using different algorithms, alternative strategies to combine embeddings from different resources, and the ICU prediction based on a variety of machine learning approaches.

The approach was applied to the MIMIC-III dataset (Johnson et al., 2016) and compared to an established baseline and a state-of-the-art approach on the 30-day ICU readmission probability at the end of the stay (Lin et al., 2019). The prediction formulation was also extended to account for predictions made throughout the ICU stay as more information is collected on a patient. The knowledge graph embeddings-based approach achieved an AUC of 0.815 (0.2 more over the state of the art and the baseline) indicating that enriching data with associated semantics as encoded in domain ontologies and knowledge graphs improved ICU readmission prediction over EHR data.

Acknowledgments: This work received financial support though a research scholarship for masters students, ref.a 716 PEI 5BI-Lic LASIGE. The work is also funded by the FCT through LASIGE Research Unit, ref. UIDB/00408/2020 and ref. UIDP/00408/2020.

References:

Correa, T., Ponzoni, C., Rabello, R., Serpa, A., Assuncao, M., Pardini, A., and Shettino, G. (2017). Readmission to intensive care unit: incidence, risk factors, resource use and outcomes: a retrospective cohort study.

Huang, L., Shea, A., Qian, H., Masurkar, A., Deng, H., and Liu, D. (2019). Patient clustering improves efficiency of federated machine learning to predict mortality and hospital stay time using distributed electronic medical records. Journal of Biomedical Informatics, 99:103291.

Johnson, A., Pollard, T., Shen, L., Lehman, L.-w., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L., and Mark, R. (2016). Mimic-iii, a freely accessible critical care database. Scientific Data, 3:160035.

Lai, J.-I., Lin, H.-Y., Lai, J., Lin, P.-C., Chang, S.-C., and Tang, G.-J. (2012). Readmission to the intensive care unit: A population-based approach. Journal of the Formosan Medical Association = Taiwan yi zhi, 111:504–9.

Lin, Y.-W., Zhou, Y., Faghri, F., Shaw, M., and Campbell, R. (2019). Analysis and prediction of unplanned intensive care unit readmission using recurrent neural networks with long short-term memory. PLOS ONE, 14:e0218942.

Scherpf, M., Gr¨aßer, F., Malberg, H., and Zaunseder, S. (2019). Predicting sepsis with a recurrent neural network using the mimic iii database. Computers in Biology and Medicine, 113:103395.

Suresh, H., Gong, J., and Guttag, J. (2018). Learning tasks for multitask learning: Hete-rogenous patient populations in the icu.

Transcriptomic effects of cigarette smoking across human tissues

Rogério Ribeiro[1,2], Raquel García-Pérez[3], José Miguel Ramiréz[3], Marta Melé[3] & Ferreira P. G.[1,2]

[1] InescTec
[2] DCC – FCUP
[3] Barcelona Supercomputing Center

Cigarette smoking is the leading cause of preventable deaths worldwide with a mortality rate of 7 million deaths per year. It constitutes a causative factor for several pulmonary diseases, such as chronic obstructive pulmonary disease and lung cancer as well as heart diseases and immune system issues. Nevertheless, most studies targeting the transcriptome of smokers have focused on airway and whole blood samples. In this work, we leveraged the transcriptome data from the GTEx consortium, along with the corresponding detailed phenotypic annotation to uncover the effect of smoking on 49 human tissues, by performing differential expression and differential splicing analysis. Our results indicate that the most affected tissue was the lung, followed by esophagus mucosa, thyroid and pancreas. Most of the differential expressed genes were tissue specific indicating that each tissue has a unique response to tobacco smoking. Remarkably, we found that CYP1A1 and AHRR, which have previously been associated with smoking, were expressed in over 20 tissues, with GPR15 being differentially expressed in 15. Since the lung samples yielded the most significant changes, we asked if the transcriptomic differences were a result of perturbation at single cell level. We employed a detailed single-cell dataset of lung samples, with 38 cell types (including rare lung cell types) to deconvolute bulk lung samples. We compared the cell populations of smokers vs non-smokers and uncovered changes in the proportions within the immune cells repertoire. Furthermore, differentially expressed analysis also revealed changes pertaining to specific cell types. Finally, since smoking cessation has been indicated to decrease the risk of diseases, we identified tissue-wide reversible and permanent changes related to tobacco smoking. Our results across different analyses revealed that smoking induces depth dysregulation of the human transcriptome, which might explain the elevated disease risk for smoker individuals.

Bioinformatics Applications

New Insights into the Catalytic Mechanism of Plastic Degrading Enzyme IsPETase: a QM/MM computational study

Rita P. Magalhães[1,2], Henrique S. Fernandes[1,2] and Sérgio F. Sousa[1,2]

[1] Associate Laboratory i4HB – Institute for Health and Bioeconomy, Faculdade de Medicina, Universidade do Porto, 4200-319 Porto, Portugal
[2] UCIBIO/REQUIMTE, BioSIM – Departamento de Medicina, Faculdade de Medicina da Universidade Do Porto, Alameda Prof. Hernâni Monteiro, 4200-319 Porto, Portugal

Plastic accumulation is one of the main environmental issues of our time. Over the years, several enzymes with the ability to hydrolize plastic have been discovered and characterised.[1] In 2016, two enzymes capable of degrading PolyEthylene Terepthtalate (PET) were discovered: *Is*PETase and *Is*MHETase, from *Ideonella sakaiensis*.[2]

In this work, the catalytic mechanism of *Is*PETase was studied by a subtractive ONIOM QM/MM methodology.[3] The system was divided in two regions: the high-level (HL) layer, calculated with density functional theory (DFT), and the low-level (LL) layer, calculated with molecular mechanics (MM). The HL layer included the catalytic residues (Ser160, Asp206, His237), the oxyanion hole, and three stabilising residues we have found to highly impact the mechanism.

The reaction was found to progress in four distinct steps, divided in two major events: formation of the first transition intermediate and hydrolysis of the adduct. The transition state and respective reactant and product of each step were fully characterised and described, and the full energy profile of the catalytic reaction was mapped out. The determined turnover rate agrees with the current experimental findings regarding kinetics. Furthermore, in this study, we have highlighted the importance of using a large QM region and clarified the role of three residues interacting with catalytic aspartate. These findings will allow for a more rational and direct enzyme design, so that catalytic efficiency can be improved.

Acknowledgements This work was supported by the Applied Molecular Biosciences Unit—UCIBIO and Associate Laboratory i4HB, which are financed by national funds from FCT (UIDP/04378/2020, UIDB/04378/2020, and LA/P/0140/2020). RM acknowledges FCT for the PhD grant 2020.09087.BD.

References

1 R. P. Magalhães, J. M. Cunha and S. F. Sousa, *Int. J. Mol. Sci.* , 2021, 22.
2 S. Yoshida, K. Hiraga, T. Takehana, I. Taniguchi, H. Yamaji, Y. Maeda, K. Toyohara, K. Miyamoto, Y. Kimura and K. Oda, *Science (80-.).*, 2016, 351, 1196–1199.
3 R. P. Magalhães, H. S. Fernandes and S. F. Sousa, *Isr. J. Chem*., 2020, 60, 655–666.

The catalytic mechanism of Pdx2 glutaminase: a QM/MM approach

André F. Pina[1,2], Sérgio F. Sousa[1,2], Nuno M. F. S. A. Cerqueira[1,2]

[1] Associate Laboratory i4HB – Institute for Health and Bioeconomy, Faculty of Medicine, University of Porto, 4200-319 Porto, Portugal
[2] UCIBIO – Applied Molecular Biosciences Unit, BioSIM – Department of Biomedicine, Faculty of Medicine, University of Porto, 4200-319 Porto, Portugal

Pdx2, the glutaminase subunit of the pyridoxal 5'-phosphate (PLP) synthase, is a key enzyme in the synthesis of PLP. It employs a non-canonical Cys-His-Glu triad to catalyse the deamination of glutamine to glutamate and ammonia – the source of the nitrogen of PLP. For this reason, Pdx2 is considered a novel and promising drug target against diseases such as Malaria and Tuberculosis, whose pathogens rely on this enzyme to obtain PLP. Therefore, the catalytic mechanism of Pdx2 was studied with atomic detail using the computational ONIOM QM/MM methodology with an 80/81 atoms QM region, which includes all catalytic relevant residues, treated at the DLPNO- CCSD(T)/CBS//B3LYP/6-31G(d,p):ff14SB level. The results demonstrate that the catalytic mechanism of Pdx2 occurs in six steps divided into four main stages: (i) activation of Cys87, (ii) deamination of glutamine with formation of the glutamyl-thioester intermediate, (iii) hydrolysis of the formed intermediate, and (iv) enzymatic turnover. The rate-limiting step of the complete catalytic mechanism is the hydrolysis of the glutamyl- thioester intermediate (18.2 kcal.mol-1), which closely agrees with the available kinetic data (19.1–19.5 kcal mol-1). The catalytic mechanism of Pdx2 differs from other known amidases in three main points: i) it requires the activation of the nucleophile Cys87 to a thiolate; ii) the hydrolysis occurs in a single step without formation of a second tetrahedral intermediate, and iii) Glu198 does not have a direct role in the catalytic process.

References

Pina, A., Sousa, S. and Cerqueira, N.M.F.S.A. (2021), The catalytic mechanism of Pdx2 glutamine driven by a Cys-His-Glu triad: a computational study. ChemBioChem.

Accepted Author Manuscript. DOI: 10.1002/cbic.202100555

Bioinformatic strategies in grapevine genomics, from biodiversity to domestication

Sara Freitas[1,2,3], Miguel Carneiro[1,2,3], Herlander Azevedo[1,2,3]

[1] CIBIO, Centro de Investigação em Biodiversidade e Recursos Genéticos, InBIO Laboratório Associado, Campus de Vairão, Universidade do Porto, 4485-661 Vairão, Portugal;
[2] Departamento de Biologia, Faculdade de Ciências, Universidade do Porto, 4099-002 Porto, Portugal;
[3] BIOPOLIS Program in Genomics, Biodiversity and Land Planning, CIBIO, Campus de Vairão, 4485-661 Vairão, Portugal.

Grapevine (*Vitis vinifera L.*) diversity richness results from a complex domestication history over multiple historical periods. Expansion of human activity led to the creation of thousands of varieties with extensive phenotypic diversity. Unfortunately, the recent favoring of specific varieties/clones, and the globalisation-driven exposure to pathogens, has led to extensive genetic erosion in this widely cultivated and economically significant crop. Fighting this genetic erosion whilst addressing issues of resilience to climate change, yield and other traits, requires a crucial understanding of the genetic basis of grapevine variation. This scientific field has witnessed significant advances due to the use of genomics approaches, enabled by Next Generation Sequencing. Here, NGS-driven whole genome resequencing strategies have been used to tackle multiple aspects associated with the extant biodiversity present in grapevine germplasm, including a clarification of different features of its recent evolutionary history that suggest a meaningful role of the Iberian Peninsula in grapevine domestication. These different aspects of grapevine biology, which require genome-level analysis, employ multiple genomics and bioinformatics strategies that help bridge the gap between population history, genomic variation and gene function.

Funding: Fundação para a Ciência e Tecnologia (FCT/MCTES) for project GrapeVision (PTDC/BIA-FBT/2389/2020) and support to H.A. CEECIND/00399/2017/CP1423/CT0004); FCT/MCTES and POCH/NORTE2020/FSE for support to S.F. (SFRH/BD/120020/2016); FCT/MCTES and POPH-QREN/FSE for support to M.C. (CEECINST/00014/2018/CP1512/CT0002).

Systems Biology

Drug Target Identification based on the construction of a Genome-scale metabolic model for the human pathogen *Candida parapsilosis*

Diogo Couceiro 1,2 , Romeu Viana 1,2, Tiago Carreiro 1,2, Oscar Dias 3 , Isabel Rocha 4 , Miguel Cacho Teixeira 1,2

1 Department of Bioengineering, Instituto Superior Técnico, University of Lisbon, Lisboa, Portugal
2 iBB - Institute for Bioengineering and Biosciences, Associate Laboratory Institute for Health and Bioeconomy - i4HB, Lisboa, Portugal
3 CEB - Centre of Biological Engineering, Universidade do Minho, Braga, Portugal
4 ITQB Nova - Instituto de Tecnologia Química e Biológica António Xavier, Lisboa, Portugal

Candida parapsilosis has seen one of the most significant rises in incidence among pathogenic *Candida spp*., often taking second place only to C. albicans [1]. Addingto this increased incidence is the rise in resistance to first line antifungals and lack ofadequate alternative therapeutics, not only for *C. parapsilosis* but throughout the genus [2,3]. Genome Scale Metabolic Models (GSMMs) have risen as a powerful in silico tool for the understanding of pathogenesis due to their systems view of metabolism and, above all, drug target predictive capacity [4–6].

In this study the first validated GSMM for *C. parapsilosis* was constructed – iDC1003 – comprising 1003 genes, 1804 reactions and 1278 metabolites, across four compartments and an intercompartment. *In silico* growth parameters as well as predicted utilisation of several metabolites as sole carbon or nitrogen sources were experimentally validated, with iDC1003 showing realiably predictive accuracy. Finally, iDC1003 was exploited as a platform for the prediction of 147 essential enzymes in mimicked host conditions, with 56 also predicted as essential in *C. albicans* and *C. glabrata*. These promising drug targets include, besides those already used as targets of clinically used antifungals, others that seem to be entirely new and worth further scrutiny. The obtained results strengthen the position of GSMMs as promising platforms for drug target discovery and as guiding tools for designing novel effective antifungal therapies.

1. Trofa, D., Gácser, A. & Nosanchuk, J. D. Candida parapsilosis, an emerging fungal pathogen. *Clin. Microbiol*. Rev. 21, 606–625 (2008).
2. Castanheira, M., Deshpande, L. M., Messer, S. A., Rhomberg, P. R. & Pfaller, M. A. Analysis of global antifungal surveillance results reveals predominance of Erg11 Y132F alteration among azole-resistant Candida parapsilosis and Candida tropicalis and country-specific isolate dissemination. *Int. J. Antimicrob. Agents* 55, 105799 (2020).
3. Silva, S. *et al*. Candida glabrata, Candida parapsilosis and Candida tropicalis: Biology, epidemiology, pathogenicity and antifungal resistance. *FEMS Microbiol. Rev.* 36, 288–305 (2012).
4. Gu, C., Kim, G. B., Kim, W. J., Kim, H. U. & Lee, S. Y. Current status and applications of genome-scale metabolic models. *Genome Biol*. 20, 1–18 (2019).
5. Raškevičius, V. *et al*. Genome scale metabolic models as tools for drug design and personalised medicine. *PLoS One* 13, 1–14 (2018).
6. Viana, R. *et al*. Genome-scale metabolic model of the human pathogen candida albicans: A promising platform for drug target prediction. *J. Fungi* 6, 1–19 (2020).

A Continuous Approach For Modeling Fermentation Phases In Wine Production

Artai Moimenta, David Henriques and Eva Balsa-Canto.

Biosystems and bioprocess engineering group, IIM-CSIC, Vigo, Spain

The yeast *Saccharomyces cerevisiae* is an essential microorganism in food biotechnology, particularly in wine and beer making. During wine fermentation, yeasts transform sugars present in the grape juice into ethanol and carbon dioxide. The process occurs in batch conditions and is an anaerobic process for the most part. Previous studies linked limited-nitrogen conditions with problematic fermentations, with negative consequences for the performance of the process and the quality of the final product. It is, therefore, of the highest interest to anticipate such problems through mathematical models.

In batch operation, cell culture follows a growth curve with the following phases: lag- phase, exponential growth, growth under nutrient limitation, stationary phase, and cellular decay. Current dynamic models of yeast metabolism explain the measured dynamics of biomass growth, carbon sources uptake, and the production of relevant primary metabolites reasonably well. Unfortunately, several aromatic metabolites relevant for wine production are associated with other fermentation phases, and therefore a better description of the process is still necessary.

Recently, we proposed a model to explain fermentation under nitrogen-limited anaerobic conditions. This model separated biomass formation into two phases: growth and carbohydrate accumulation. Growth was modeled using the well-known Monod equation, while carbohydrate accumulation was modeled by an empirical function analogous to a proportional controller activated by the limitation of available nitrogen.

In this work we extend our previous fermentation model with kinetic rates for aroma formation. These rates are associated with the secondary growth-phase mechanism. The final model was used to successfully explain biomass production, main extracellular metabolites (glucose, ethanol, glycerol, etc.), and aromas.

MEWpy: Modelled by Man, designed by Nature

Vítor Pereira

Centro de Engenharia Biológica, Departamento de Informática, Universidade do Minho

Finding genetic modifications to build biofactories, analysing microbial communities and host- pathogen interactions, or studying the evolution of cross-feeding interactions between organisms continues to be significant Systems Biology challenges. Beyond understanding the inter and intracellular interactions, identifying optimised genetic modifications or communities compositions can open the door to new environmental-friendly solutions and disruptive disease therapeutics. We have been combining constraint-based modelling formulations with artificial intelligence for years. Interestingly, bio-inspired optimisation procedures have been successfully applied to bioinformatics problems, showcasing that Nature offers the means, the inspiration and the solutions. MEWpy follows such a path and provides tools to design new strains, optimise microbial communities and integrate omics data, while SDDB makes stain designs available to the community.

Bioinformatics approaches for engineering L-Tyrosine production in Escherichia coli

Maria João Lopes, Joana L. Rodrigues and Oscar Dias

Centre of Biological Engineering, University of Minho, Portugal

Aromatic compounds, such as aromatic amino acids, petrochemical aromatics, and aromatic polymers, are very important industrial materials. Among them, *L*-tyrosine (*L*-Tyr) has been considered an appealing metabolite to produce thanks to its wide variety of applications in the pharmaceutical and chemical industries. This metabolite is an important precursor of a diverse number of secondary metabolites or natural products, such as phenylpropanoic acids, flavonoids, curcuminoids, stilbenoids, Parkinson's disease drug 3,4-dihydroxy-L-phenylalanine (L-DOPA) and melanin [1]–[4]. There are several studies where *Escherichia coli* *L*-Tyr overproducing strains are used, but there is still a need to increase its yield to make the bioprocess economically feasible [1], [2].

Numerous computational tools have been developed for strain design that identify genetic modification strategies that increase targeted biochemicals production [5]. The construction of a biochemical network map, and the development of mathematical models and simulations that reproduce experimental data and phenotypes under different conditions, allow computer-aided cell design [6]. Stoichiometric models alone cannot quantitatively measure the effect of concentration levels, enzyme saturation and regulation. Kinetic models yield a system of ordinary differential equations that describe the time evolution of metabolite concentrations, enzyme activities and reaction fluxes, therefore have the potential to capture these interdependencies [5], [7]. These types of models may be used to model *E. coli* to overproduce *L*-Tyr.

The aim of this work was to perform the *in silico* insertion of *L*-Tyr pathway in a kinetic model of the central carbon metabolism of *E. coli*, which will allow to identify the best target genes for the design of the *L*-Tyr overproducing strain. Three stoichiometric models, iML1515 [8], iJO1366 [9] and iJR904 [10], as well as three kinetic models, the Millard [11], the Oliveira [12] and the Jahan [6], were used to design an accurate model containing all the reactions needed to produce *L*-Tyr. This new model was then used to test the different knock-in/out strategies to design an *E. coli* *L*-Tyr overproducing strain.

References

[1] D. Juminaga *et al.*, "Modular engineering of L-tyrosine production in Escherichia coli," *Appl. Environ. Microbiol.*, vol. 78, no. 1, pp. 89–98, 2012, doi: 10.1128/AEM.06017-11.

[2] B. Kim, R. Binkley, H. U. Kim, and S. Y. Lee, "Metabolic engineering of Escherichia coli for the enhanced production of l-tyrosine," *Biotechnol. Bioeng.*, vol. 115, no. 10, pp. 2554–2564, 2018, doi: 10.1002/bit.26797.

[3] G. Hernández-Chávez, A. Martínez, and G. Gosset, "Metabolic engineering strategies for caffeic acid production in Escherichia coli," *Electron. J. Biotechnol.*, vol. 38, pp. 19–26, 2019, doi: 10.1016/j.ejbt.2018.12.004.

[4] J. L. Rodrigues, K. L. J. Prather, L. D. Kluskens, and L. R. Rodrigues, "Heterologous Production of Curcuminoids," *Microbiol. Mol. Biol. Rev.*, vol. 79, no. 1, pp. 39–60, 2015, doi: 10.1128/mmbr.00031-14.

[5] A. Khodayari, A. R. Zomorrodi, J. C. Liao, and C. D. Maranas, "A kinetic model of Escherichia coli core metabolism satisfying multiple sets of mutant flux data," *Metab. Eng.*, vol. 25, pp. 50–62, 2014, doi: 10.1016/j.ymben.2014.05.014.

[6] N. Jahan, K. Maeda, Y. Matsuoka, Y. Sugimoto, and H. Kurata, "Development of an accurate kinetic model for the central carbon metabolism of Escherichia coli," *Microb. Cell Fact.*, vol. 15, no. 1, pp. 1–19, 2016, doi: 10.1186/s12934-016-0511- x.

[7] A. Khodayari, A. Chowdhury, and C. D. Maranas, "Succinate Overproduction: A Case Study of Computational Strain Design Using a Comprehensive Escherichia coli Kinetic Model," *Front. Bioeng. Biotechnol.*, vol. 2, no. January, 2015, doi: 10.3389/fbioe.2014.00076.

[8] J. M. Monk et al., "iML1515, a knowledgebase that computes Escherichia coli traits," *Nat. Biotechnol.*, vol. 35, no. 10, pp. 904–908, Oct. 2017, doi: 10.1038/nbt.3956.

[9] J. D. Orth et al., “A comprehensive genome-scale reconstruction of Escherichia coli metabolism-2011,” *Mol. Syst. Biol.*, vol. 7, no. 535, pp. 1–9, 2011, doi: 10.1038/msb.2011.65.
[10] J. L. Reed, T. D. Vo, C. H. Schilling, and B. O. Palsson, “An expanded genome- scale model of Escherichia coli K-12 (iJR904 GSM/GPR).,” *Genome Biol.*, vol. 4, no. 9, pp. 1–12, 2003, doi: 10.1186/gb-2003-4-9-r54.
[11] P. Millard, K. Smallbone, and P. Mendes, “Metabolic regulation is sufficient for global and robust coordination of glucose uptake, catabolism, energy production and growth in Escherichia coli,” *PLoS Comput. Biol.*, vol. 13, no. 2, pp. 1–24, 2017, doi: 10.1371/journal.pcbi.1005396.
[12] A. Oliveira, J. Rodrigues, E. C. Ferreira, L. Rodrigues, and O. Dias, “A kinetic model of the central carbon metabolism for acrylic acid production in Escherichia coli,” *PLOS Comput. Biol.*, vol. 17, no. 3, p. e1008704, Mar. 2021, doi: 10.1371/journal.pcbi.1008704.

COAST: A bioinformatic tool to identify the closest proteomes

Diogo Macedo (1), Sílvio Santos (2), Óscar Dias (2)

(1) Departamento de Informática, Universidade do Minho
(2) Centro de Engenharia Biológica, Universidade do Minho

Recent advances in sequencing technologies, DNA manipulation and synthetic biology approaches, together with their potential applications in many fields of research, led to an increasing amount of DNA sequence and whole-genome data contributing to the accumulation of enormous raw data sets. Such massive data *per se* is meaningless and requires the development of efficient bioinformatics tools to analyse and mine data from it, constituting a major challenge, mainly if we want to look globally at whole genomes.

Identifying homologous genes and analysing their synteny are essential to understand the rules of genome structure and predict gene or protein structure and function. Besides the growing development of bioinformatics, tools able to perform comparative analysis between whole genomes is still scarce, and the existing ones rely on a manual, awkward and slow process. Moreover, there is a lack of tools that identify the closest genomes and/or proteomes from discrete sequence datasets, a gap that needs to be fulfilled.

We developed COAST, The Comparative Omics Alignment Search Tool, considering all this knowledge. COAST is a command-line tool that exploits existing alignment algorithms and databases to provide a simple search tool capable of identifying organism with similar proteomes or taxonomy closeness. The algorithm scores closeness using scores based on the Average Amino acid Identity (AAI), a well- recognised and used relatedness score.(1) These organisms will be required for further downstream comparative analysis, improving the annotation of new genomes by identifying and characterising the functions for newly found putative genes based on the homologs, leading to a deeper understanding of one organism in the context of another and giving insights into the structure and function of genes and genomes aiding on the taxonomic and phylogenetic studies of new organisms. The tool is available for the Galaxy platform (2,3), through a simple webapp that can be integrated into larger Galaxy workflows.

References:
1. Konstantinidis KT, Tiedje JM. Towards a Genome-Based Taxonomy for Prokaryotes. J Bacteriol. 2005 Sep 15;187(18):6258–64.
2. Afgan E, Baker D, Batut B, van den Beek M, Bouvier D, Čech M, et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. Nucleic Acids Research. 2018 Jul 2;46(W1):W537–44.
3. Blankenberg D, Von Kuster G, Bouvier E, Baker D, Afgan E, Stoler N, et al. Dissemination of scientific software with Galaxy ToolShed. Genome Biology. 2014 Feb 20;15(2):403.

DeepSweet: development of Deep Learning models to predict sweetness

João Capela[1], João Correia[1], Vítor Pereira[1], Miguel Rocha[1]

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

The demand for conventionally used sugars decreased due to related health issues in the past few years. On the contrary, the non-caloric sweeteners market is increasing significantly, providing incentives for their production. Hence, developing innovative strategies for designing non-toxic molecules with a sweet taste is valuable for the food industry. Accordingly, Machine Learning (ML) models became relevant to conceive molecular Structure-Activity Relationships (SAR), having been used recently to predict the flavour of molecules.

Herein, we deliver the largest known dataset with sweet molecules, containing 9541. This dataset allowed us to implement standard Machine and Deep Learning (DL) pipelines with DeepMol (https://github.com/BioSystemsUM/DeepMol). DeepMol is a python package that provides a smoother approach to ML/DL pipelines applied to chemoinformatics. The package covers the molecules' preprocessing, generation of features, selection, model construction, hyperparameter optimisation and feature explainability. As for the model construction, DeepMol uses Tensorflow (https://www.tensorflow.org/), Keras (https://keras.io/), Scikit-learn (https://scikit-learn.org/) and DeepChem (https://deepchem.io/) to build custom ML/DL models or use pre-built ones.

We showcase that Textual Convolutional Neural Networks (TextCNN), Graph Convolutional Networks (GCN), and Deep Neural Networks (DNNs) outperformed most of the traditional "shallow" learning approaches. Accordingly, we present the first DL platforms to predict sweetness, providing means to repurpose existing molecules and guide the design of novel sweeteners.

Herein, we evaluated sixty million compounds from PubChem using these models. After this evaluation, we delivered a dataset of 67724 compounds that present high probabilities of being sweet. Quick searches in aspartame and guanidine derivatives on this set revealed that at least 13 molecules were sweet. This evaluation corroborates the usefulness of our approach to find new sweeteners, valuable to expand food chemistry databases.

Finally, we performed a SHapley Additive exPlanations analysis on one of the best models to assess the most important features for predicting molecules as sweet. We showed that this model is in accordance with the most accepted theories on the molecular basis of sweetness.

This work delivers the largest known dataset of sweeteners and the first Deep Learning-based platform to predict sweetness. The whole pipeline and data are available at https://github.com/BioSystemsUM/DeepSweet.

Evolution of multiheme cytochromes involved in diverse biogeochemical cycles

Ricardo Soares[1,2], Nazua L. Costa[1], Catarina M. Paquete[1], Claudia Andreini[3] and Ricardo O. Louro[1]

1 Instituto de Tecnologia Química e Biológica António Xavier, Universidade Nova de Lisboa, Portugal
2 Instituto Nacional de Investigação Agrária e Veterinária, Portugal
3 Magnetic Resonance Center and Department of Chemistry, University of Florence, Sesto Fiorentino, Italy

Prokaryotes harboring multiheme *c*-type cytochromes (MHC) catalyse focal chemical reactions within diverse biogeochemical cycles of the elements that occur on earth. MHC are proteins that contain multiple heme cofactors covalently bound to the polypeptide chain via specific binding motifs and often perform the final reduction step of various anaerobic respiratory metabolisms. The origin of these MHC dates back to the Archaean Eon where the atmosphere and oceans were not yet oxygen-rich environments. Some extant prokaryotes still conserve these ancient metabolic capacities that account for a high metabolic versatility and/or the colonisation of specific anaerobic habitats.

MHC involved in the nitrogen, sulfur and iron biogeochemical cycles are homologous based on sequence and structural similarities. Nevertheless, their origin and evolution remain elusive due to the lack of studies that are able to perform a systematic investigation of their evolutionary relationships. The ubiquitous pentaheme nitrite reductase (NrfA) was proposed to be at the origin of octaheme cytochromes (Occs) involved in the nitrogen and sulfur biogeochemical cycles. However, the mechanism by which NrfA evolve to an Occ has never been established. Recently, the structure of the nonaheme cell-surface OcwA from the Gram- positive *Thermincola potens* JR showed homology with NrfA at its C-terminus and with cytochrome c_{554} at its N-terminus. Based on this observation, an NrfA ancestor was proposed to fuse with a cytochrome c_{554} ancestor to originate nonaheme cytochromes similar to OcwA. Since OcwA harbors both catalytic hemes of NrfA and cytochrome c_{554}, while the different Occs only harbor one of these, loss of either catalytic heme of OcwA was proposed to be the mechanism to give rise to extant Occs.

Here, we have searched for all possible homologous MHC involved in the nitrogen, sulfur and iron biogeochemical cycles. By performing a phylogenetic analysis using character and structure data collected from different biological public databanks, we revealed a different storyline from the previous proposals. Here, an alternative proposal for the evolution of these MHC and its implications are presented. Based on these results, the evolution of diverse biogeochemical cycles are discussed, allowing to propose a last common ancestral of these MHC.

Explainable AI for Understanding Associations Between Disease-Related Genes

José Zenóglio de Oliveira[1], Francisco Pinto[2], Cátia Pesquita[1]

[1] LASIGE, Faculdade de Ciências, Universidade de Lisboa.
[2] BioISI, Faculdade de Ciências, Universidade de Lisboa.

In recent years, we have witnessed the impressive successes of “black-box” models such as deep neural networks, and these have contributed to unravel new Artificial Intelligence (AI) and Machine Learning (ML) applications with the purpose of growing scientific knowledge. However, in areas such as genomics, drug-discovery, and pathology analysis, these applications are limited in the sense they fail to provide human-understandable logical decisions and often lack some degree of explainability[1].

One well-known strategy for understanding molecular and pathological mechanisms is discovering disease-associated genes (DGs), which may provide new candidate genes for targeted therapy and research tracks for developing new drugs. However, in prioritising genes for further investigation, the need for AI and ML models to provide human understandable explanations becomes ever more apparent[2].

This is the case for a Network-Based approach called S2B[3] (Specific-Double Betweenness), which uses protein-protein interaction (PPI) networks to predict DGs simultaneously associated with two diseases. S2B has shown promising results but it only focuses on PPI networks to provide DGs, and even though we might generate interesting candidates with S2B, not only is the interactome itself limited, but we are still lacking contextual information about the generated candidate genes.

We believe that Ontologies may be the solution for providing more scientific knowledge to this method, as well as the explainability it requires. We developed a novel approach to enrich PPI networks with information from ontologies and assessed the impact of using different methods of filtering the content provided to the network to obtain the best possible results with S2B. S2B’s performance in predicting genes associated with Amyotrophic Lateral Sclerosis (ALS) and Spinal Muscular Atrophy (SMA) was evaluated with different PPI networks as an input, some of them containing information from the Gene Ontology[4]. Preliminary results show an increase in S2B performance with the use of a network containing only physical PPIs and the Gene Ontology, and these results motivate further work in refining a method to retain important information and enriching PPI networks with ontologies.

Acknowledgements: The work is funded by the FCT through LASIGE Research Unit, ref. UIDB/00408/2020 and ref. UIDP/00408/2020.) and partially supported by UIDB/04046/2020 and UIDP/04046/2020 Centre grants from FCT, Portugal (to BioISI).

References

1. Goebel R, Chander A, Holzinger K, *et al*. Explainable AI : the new 42 ? *2nd Int Cross-Domain Conf Mach Learn Knowl- edge Extr*. Published online 2018.
2. Doran D, Schulz S, Besold TR. What Does Explainable AI Really Mean? A New Conceptualization of Perspectives. Published online 2017. Accessed December 1,2021. http://amueller.github.
3. Garcia-Vaquero ML, Gama-Carvalho M, Rivas JD Las, Pinto FR. Searching the overlap between network modules with specific betweeness (S2B) and its application to cross-disease analysis. *Sci Rep*. 2018;8(1):1-10. doi:10.1038/s41598-018-29990-7

4. Gene Ontology Consortium T, Ashburner M, Ball CA, et al. Gene Ontology: tool for the unification of biology NIH Public Access Author Manuscript. *Nat Genet*. 2000;25(1):25-29. doi:10.1038/75556

Feasibility of applying shotgun metagenomic analyses to grapevine leaf, rhizosphere and soil microbiome characterization

David Azevedo-Silva[1,2,3], Jacob Agerbo Rasmussen[4], Miguel Carneiro[1,2,3], M. Thomas P. Gilbert[4,5], Herlander Azevedo[1,2,3]

[1] CIBIO, Centro de Investigação em Biodiversidade e Recursos Genéticos, InBIO Laboratório Associado, Campus de Vairão, Universidade do Porto, 4485-661 Vairão, Portugal;
[2] Departamento de Biologia, Faculdade de Ciências, Universidade do Porto, 4099-002 Porto, Portugal;
[3] BIOPOLIS Program in Genomics, Biodiversity and Land Planning, CIBIO, Campus de Vairão, 4485-661 Vairão, Portugal;
[4] Center for Evolutionary Hologenomics, The GLOBE Institute, University of Copenhagen, 1353 Copenhagen, Denmark;
[5] University Museum, Norwegian University of Science and Technology, Trondheim, Norway.

The characterisation of plant microbiomes using metabarcoding strategies is expected to be progressively replaced by shotgun metagenomic (SMg) approaches. In the present report we explored the potential of applying SMg to grapevine leaf, rhizosphere and soil samples. Our strategy involved combining a single method for column-based DNA extraction of multiple tissues, with sequencing using the BGISEQ short-read NGS platform. This study entailed DNA isolation, library construction, sequencing and early bioinformatics treatment of read data, including the use of subsampling as a proxy for detection of sample microbial proportions. The combination of an innovative sequencing platform with recent mapping tools allowed for the characterisation of the microbiome associated with the grapevine leaf, rhizosphere and adjacent soil. We coupled a robust single extraction/library preparation protocol suitable for contrasting samples, with cost- effective BGISEQ short-read NGS technology, to generate a landscape of grapevine- associated microbial diversity. Shotgun metagenomics is emerging as a state-of-the-art approach to grapevine microbiome characterisation. The present report provides a detailed and technical roadmap targeting multiple aspects of this important approach.

Funding: Fundação para a Ciência e Tecnologia (FCT/MCTES) for support to H.A. (CEECIND/00399/2017/CP1423/CT0004); FCT/MCTES and POPH-QREN/FSE for support to M.C. (CEECINST/00014/2018/CP1512/CT0002); Danish National Research Foundation award DNRF143 to M.T.P.G.

Improving data mining of PPI networks by combining deep learning methods with knowledge graphs

Laura Balbi, Cátia Pesquita lbalbi@lasige.di.fc.ul.pt

LASIGE, Faculdade de Ciências da Universidade Lisboa

Many bioinformatics problems pertain to large, highly complex amounts of biological data, that are often modelled in a graph-like arrangement to allow for a systemic-level analysis of data. A graph provides a data structure for knowledge representation, useful for the description and analysis of relationships in all kinds of knowledge domains, including biological systems[1]. It can model a multitude of biological processes, like transportation and metabolic processes, and capture associations between any type of biological entity – e.g. between proteins, by building up a Protein-Protein Interaction Network.

Several approaches have been made to apply machine learning (ML) models to data represented on graphs[2], either through the use of dense low-dimension vector representations of the different graph elements, or by building the models as representation-learning approaches that can receive and analyse such graph-structured data – prompting graph neural networks [3,4].

A form of knowledge representation that allows for the conceptualisation and specification of domains of interest is by means of the use of ontologies[5]. With the increased use of biomedical ontologies[6] for representing and structuring the existing biological knowledge by their meaning and relationships, large volumes of biological entities have been represented and organised into Knowledge Graphs (KG) [7].

Usually, for machine learning models to be employed in prediction tasks over knowledge graphs, their learning of the KG's components' semantic information is dependent on it being depicted through vector-based representations that can be processed by the models. However, when both biological data and biological knowledge are represented as graphs, an opportunity to directly enrich the data graph with knowledge about its entities arises [8] . Thus, we aim to explore how complementing a data graph with knowledge of its biological data may allow different graph mining approaches to leverage the additional information to improve their predictive performance, in particular over PPI networks.

Acknowledgements: This work is funded by the FCT through the LASIGE Research Unit, ref.UIDB/00408/2020 and ref. UIDP/00408/2020.

References

[1] – Georgios A. Pavlopoulos, Maria Secrier, Charalampos N. Moschopoulos, Theodoros G. Soldatos, Sophia Kossida, Jan Aerts, Reinhard Schneider, & Pantelis G. Bagos. 2011. Using graph theory to analyse biological networks. BioData Mining, 4(1), 1–27. https://doi.org/10.1186/1756-0381-4-10

[2] – William L. Hamilton, Rex Ying & Jure Leskovec. 2017. Inductive representation learning on large graphs. Advances in Neural Information Processing Systems, 2017-December(Nips), 1025–1035. https://dl.acm.org/doi/10.5555/3294771.3294869

[3] – Marco Gori, Gabriele Monfardini & Franco Scarselli. 2005. A new model for learning in graph domains. Proceedings. 2005 IEEE International Joint Conference on Neural Networks. 2, (2005), 729-734.

[4] – Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner & Gabriele Monfardini. 2008. The Graph Neural Network Model. IEEE Transactions on Neural Networks, 20, 1 (2009), 61-80. https://doi.org/10.1109/TNN.2008.2005605

[5] – Manuel de Sousa Ribeiro & João Leite. 2021. Aligning Artificial Neural Networks and Ontologies towards Explainable AI. Proceedings of the AAAI Conference on Artificial Intelligence. 35 (6), 4932-4940.

[6] – Darren A. Natale, Cecilia N. Arighi, Winona C. Barker, Judith Blake, Ti-Cheng Chang, Zhangzhi Hu, Hongfang Liu, Barry Smith, & Cathy H. Wu. 2007. Framework for a protein ontology. BMC bioinformatics 8, (2007), S1. https://doi.org/10.1186/1471-2105-8-S9-S1
[7] – Chang Su, Jie Tong, Yonjun Zhu, Peng Cui & Fei Wang. 2018. Network embedding in biomedical data science. Briefings in Bioinformatics, 21 (1), 182–197. https://doi.org/10.1093/bib/bby117
[8] – Xiaojun Chen, Shengbin Jia, Yang Xiang. 2020. A review: Knowledge reasoning over knowledge graph. Expert Systems with Applications, Volume 141, 112948. https://doi.org/10.1016/j.eswa.2019.112948.

Is the Crabtree effect regulated at the transcriptional level? Insights from a multi-omics model

David Henriques, Artai Rodríguez-Moimenta, Eva Balsa-Canto

Biosystems and bioprocess engineering group, IIM-CSIC, Vigo, Spain

Saccharomyces cerevisiae is a facultative anaerobe; thus, it can grow through alcoholic fermentation without oxygen. Additionally, at high external glucose, even in aerobic conditions, *S. cerevisiae* produces ethanol rather than producing biomass via the tricarboxylic acid (TCA) cycle, the most efficient process. This phenomenon is known as the Crabtree effect, and it is also observed in fast-growing bacteria (regarded as overflow metabolism) and cancer cells (regarded as the Warburg effect). Ethanol produced can later be used as a substrate when glucose is depleted; this diauxic growth pattern is sometimes referred to as the make-accumulate-consume strategy.

High glucose availability triggers the Crabtree effect in both chemostat and batch experiments of *S. cerevisiae*. While this effect is known since the early 20th century, there is still a certain debate on its activation: is it activated by a limitation in the respiratory capacity, the onset of glucose repression of the respiratory metabolism, or an overflow metabolism at the pyruvate branch point?.

At first glance, experiments showing little regulation of glycolytic enzyme expression would point towards the post-translational hypotheses. However, as quantitative analyses evolve, it is becoming more apparent that fast-growing microbes sacrifice yield (efficiency) for growth rate because the machineries for growth and efficient energy generation are costly and at odds with each other. In particular, for the yeast case, trade- offs between these two magnitudes appear to be well explained by the costs of ATP synthase and ribosomes.

As part of the research on the subject various recent works addressed quantitative multi-omics studies of *S. cerevisie* Crabtree effect, reporting detailed proteomics, exometabolomics, and biomass data. Here we propose integrating these data into a novel dynamic model to address the role of regulation in the Crabtree effect. To do so, we accounted for the relevant players and mechanisms, formulated a logic-based dynamic model, and estimated the model parameters through a multi experiment parameter estimation. Remarkably, the model considers ribosome content to explain dynamic growth, which proved to be critical for model success.

The proposed model nicely recovers all data and agrees with the hypothesis that the Crabtree effect in S. cerevisiae is controlled by glucose repression at the transcriptional level. Although the expression of the glycolytic enzyme is not affected by glucose repression, given that ATP inhibits several glycolytic enzymes, the glycolytic fluxes can be indirectly modulated by the expression of genes associated with ATP production (fermentation and oxidative phosphorylation) and demand (ribosomes, amino acid synthesis, etc.).

PhageDPO: Phage Depolymerase Finder

José Duarte pg40958@uminho.pt

Antibiotic resistance is a severe public health problem. New resistance mechanisms are rapidly emerging and spreading globally, threatening our ability to treat infections. The bacteriophages (phages) arise as a possible solution through their capability of infecting and killing bacteria. Phages are natural bacterial predators: they encode an arsenal of specialised proteins to target their bacterial hosts. One emerging protein is Phages Depolymerases (DPOs), responsible for selective recognition and degradation of bacterial cell surface decorating polysaccharides, turning the bacteria susceptible to external agents. Due to the difficulty in locating these enzymes in the phage genome, we developed PhageDPO, a DPO prediction tool, through machine learning methods.

Several classifiers were created, using different datasets and algorithms and tested through cross-validation. The datasets were composed of protein sequences retrieved from the NCBI protein database and by a different number of negative cases. Two models were selected for integration in the tool: the Support Vector Machine (SVM) model created with a dataset containing data of 4311 sequences and the Artificial Neural Network (ANN) model created with a dataset containing data of 7185 sequences. On an independent validation dataset, the SVM model presented 95% accuracy, 98% precision and 91% recall and the ANN model presented 98% accuracy, 99% precision and 96% recall. While the high precision and PECC of the SVM focus on predicting true DPO sequences and avoiding false positives, the ANN ensures that all DPOs are identified due to its high recall. PhageDPO was successfully tested in predicting DPOs of, previously characterised, phages.

PhageDPO was integrated into the Galaxy framework (https://bit.ly/3dOam2u), providing a user-friendly graphical interface for wet-lab researchers without computational skills.

PHASEfilter - filtering heterozygous variants from phased genomes

Ana Rita Guimarães[1], Ana Rita Bezerra[1], Andreia Reis[1], Gabriela Moura[1], Manuel A.S. Santos[1], Miguel Pinheiro[1]

[1] Institute of Biomedicine—iBiMED, Department of Medical Sciences, University of Aveiro, 3810-193 Aveiro, Portugal

Some diploid species present the reference genome in a phase form. With phased genomes it can significantly enhance the sensitivity and specificity of allele-specific expression measurements by enabling cross-validation of variants across multiple polymorphic sites. With this information it can also reduce the number of variations leaving the research to focus their attention for the real variations.

Based on previous assumptions we present a software that can identify heterozygosity positions between two phased references. The software starts by aligning pairs of diploid chromosomes, based on Minimap2 [1] summing the match positions and obtaining the percentage of effective alignment. With synchronisation done it is possible to identify the position of a variation, in both elements of a pair of chromosomes, allowing variants removal if it meets a following established criteria.

To classify variants it is necessary to pass two VCF files, one for each reference phase. After that, the PHASEfilter will go through the variants called in reference A and check if there are any homologous in the variants called in reference B. For each variant called in the reference A it can happen three situations: 1) both references, for the position in analysis, are equal and the variant is valid; 2) position is heterozygous in the reference and the variant reflects it, so the variant is removed; 3) position is heterozygous in the reference and the variant is homozygous, so the variant is going to “loss of heterozygosity” output.

We applied this algorithm at three different sequencing runs of Candida Albicans, a known diploid organism, and successfully filter ~70% of SNPs and ~30% INDELs. It was possible to identify the loss of heterozygosity (LOH), in both pairs of chromosomes, improving evolution analysis.

PHASEfilter provides an easy way to detect and filter heterozygous SNPs and INDELs across diploid species. It will certainly strengthen the detection of genetic changes cleaning variants that are creating noise in the data leaving the research to focus on the real variations.

Download: https://github.com/ibigen/PHASEfilter

Acknowledgements: This work was supported by FEDER (Fundo Europeu de Desenvolvimento Regional) funds through the COMPETE 2020, Operational Programme for Competitiveness and Internationalization (POCI), and by Portuguese national funds via Fundação para a Ciência e a Tecnologia, I.P. (FCT) under the projects GenomePT (POCI-01-0145-FEDER-022184), PTDC/BIA-MIB/31238/2017, PTDC/BIA-MIC/31849/2017 and PTDC/BIA-MIC/1141/2021. The iBiMED research unit is supported by FCT funds under UIDP/04501/2020. ARG is supported directly by a FCT grant (SFRH/BD/121358/2016 and COVID/BD/151731/2021).

References

1. Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics, 34:3094-3100. doi:10.1093/bioinformatics/bty191;

Predicting new therapeutic candidates for triple negative breast cancer: an ancestry-guided bioinformatic analysis

Ricardo Pinto[1,2], Bruno Cavadas[1,2], Lúcio Lara Santos[3,4], Luisa Pereira[1,2]

[1] i3S - Institute for Research & Innovation in Health, University of Porto, Porto,Portugal.
[2] IPATIMUP - Institute of Molecular Pathology and Immunology of the University ofPorto, Portugal.
[3] Experimental Pathology and Therapeutics Group and Department of Oncology, Portuguese Institute of Oncology of Porto (IPO-Porto), Portugal.
[4] ONCOCIR - Education and Care in Oncology, PALOP - Lusophone Africa, Porto, Portugal.

OBJECTIVE: To evaluate if the European and African-ancestral background impacts sensitivity to chemotherapeutic drugs in triple-negative breast cancer (TNBC) cell lines.

METHODS: Drug sensitivity data from more than 190 chemotherapeutic drugs were mined from GDSC database for 23 TNBC cell lines, attending to their European and African ancestries (Wilcoxon-Mann–Whitney statistical test). RNA-Seq raw gene expression data for TNBC cell lines were retrieved from the CCLE database, and the differential expression (using DESeq2) of 289 ADME genes was evaluated between the two ancestries, using the two-sided t-test for unpaired samples.

RESULTS: 18 compounds (Acetalax, Carmustine, Dihydrorotenone, Dinaciclib, Fulvestrant, Mirin, Pevonedistat, Sapitinib, Ulixertinib, and other nine undergoing clinical development) showed significant differential interaction between African and European TNBC cell lines, with the African ones being always more sensitive than the European. Relatively to the ADME genes, seven showed differential expression between both ancestral contexts, six of these *(SLC6A6, ABCG2, CYP3A43, CYP19A1, CYP2R1, ABCC1*) presenting a lower expression in African cell lines.

CONCLUSIONS: The aggressive TNBC is characterised by the absence of known druggable molecular targets, leading to high risk of disease recurrence and, hence, a shorter overall survival of these patients. This report showed that there are some drugs that may be effective in treating TNBC, and that most of these could in fact be even more effective in the African than in the European ancestry. A possible explanation for this is that some differentially expressed ADME genes are involved in mechanisms of drug resistance, thus the lower expression of these genes in the African TNBC cell lines decreases the resistance, and hence promotes its higher sensitivity to these drugs. These findings must be validated *in vitro* and *in vivo* further investigation.

SARS-CoV-2 variant prediction using SNVs (single nucleotide variants) in infected patient sequences

Rafael Oliveira[1], Vitor Borges[2], Gabriela Moura[1], Miguel Pinheiro[1]

[1] Institute of Biomedicine – iBiMED, Department of Medical Sciences, University of Aveiro,3810-193, Aveiro, Portugal
[2] Bioinformatics Unit, Department of Infectious Diseases, National Institute of Health, Av. Padre Cruz, 1649-016, Lisbon, Portugal

The Coronavirus (Covid-19) outbreak has ignited the scientific community in ways that no other outbreak has before. It has long been known that viruses are constantly adapting through mutations which are responsible for the different variants, some of which spread faster than others. These small mutations are the backbone behind the genetic signatures that characterise each variant. The main goal of this work is to count the SNVs (single nucleotide variants), deletions and insertions, present in SARS-CoV-2 sequences, from infected patients, and use bioinformatic and statistical tools to predict the lineages. These counts are independent from this location on the complete genome of SARS-CoV-2. The sequences (~20.000) were obtained using the GISAID[1] database, filtered by Portugal country, and were further processed with Nextstrain[2] to obtain the count of each variant in every sequence against Wuhan reference (MN908947). Due to the high dimensionality of the data (251 variants and 7 dependent outcomes/lineages) a cleaning process was made, removing variants with less than 200 counts in all datasets. First, we used an exploratory analysis technique such as a principal component analysis (PCA). To further this analysis, two multinomial logistic regression models were trained; i) a complete model with a total of 19 SNVs (substitutions and deletions) ii) a second model with five dimensions (explaining ~ 90% of the original data) from the initial PCA analysis. Although the complete model achieved higher accuracy (0.9913), some of the covariates had high correlation (>0.70). Because of this high correlation, the use of principal components could ameliorate this effect since principal component analysis joins highly correlated variables into non-correlated features/principal components. However, using the latter model would not allow for an interpretation of the coefficients as we would be using principal components instead. Last, but not least, it is important to note that even if the complete model was able to achieve high accuracy, there is a presence of separation (perfect prediction or monotone likelihood) which happens when the outcome variable separates a predictor variable completely. With the approach of using a multinomial logistic regression model it is possible to identify about 99% of lineages even losing the information of the variants positions in the genome.

RPubs: https://rpubs.com/RafaelOliveira/LogRpca

References
1. Khare, S., et al (2021) GISAID's Role in Pandemic Response. China CDC Weekly, 3(49): 1049-1051. doi: 10.46234/ccdcw2021.255
2. Hadfield et al., Nextstrain: real-time tracking of pathogen evolution , Bioinformatics (2018), doi: 10.1093/bioinformatics/bty407

To be or to be NOT: The Impact of Negative Annotation in Biomedical Semantic Similarity

Lina Andreia Gama Aveiro

Faculdade de Ciências, Universidade de Lisboa

Classical Semantic Similarity Measures don't consider negative annotations in similarity computation, and the impact that these annotations can have in this data mining technique is not well studied. As such, this work aims to understand how the addition of negative annotations impacts semantic similarity. To do so, two pairwise similarity measures, BMA and Resnik, were adapted to create the polar measures PolarBMA and PolarResnik. These were evaluated in two currently relevant scopes: protein-protein interaction prediction and disease prediction against the original measures. Pairs of proteins where the proteins were known to interact or not were taken from STRING and enriched with positive and negative annotations from the Gene Ontology. Synthetic patients were created as sets of annotations taken from the Mendelian diseases they were designed to have, as well as possible noise or imprecise annotations. Then semantic similarity was computed with both polar and non-polar measures between proteins in pairs and between patients and candidate diseases including the Mendelian diseases, as well as random diseases taken from the Human Phenotype Ontology.

To evaluate if the polar measures performed well in comparison to the baseline, a ranking according to semantic similarity was made for each measure and scope for evaluation and the rank cumulative frequencies were plotted. In PPI prediction, polar measures had an increased performance in the Molecular Function branch. In the disease prediction scope, polar measures had an improved performance of approximately ten percent. This improvement was verified in all disease prediction experiments, even with the addition of noise and imprecision. Considering the results obtained, this work concludes that negative annotations have an impact on semantic similarity, but the amplitude of this impact requires further study.

Visualisation of Knowledge Graphs for Explainable AI

Filipa Serrano, Cátia Pesquita fserrano@lasige.di.fc.ul.pt

LASIGE, Faculdade de Ciências, Universidade de Lisboa

Artificial intelligence (AI) and machine learning (ML) have been achieving great results in the biomedical domain[1]. There are valuable deep learning models, with promising results but they are "black-box". Their lack of explainability severely limits their trustability, specially in such a sensitive field, where errors in predictions can have very damaging outcomes[2,3]. Additionally, legal aspects limiting the application of black-box models are also becoming stricter[4,5].

Semantic Explanations emerge as a strategy for explainable AI applications[6]. Ontologies and Knowledge Graphs (KG) are used to model data and represent it in a connected structure, that can be used to generate semantic explanations for each patient's predicted treatment[7].

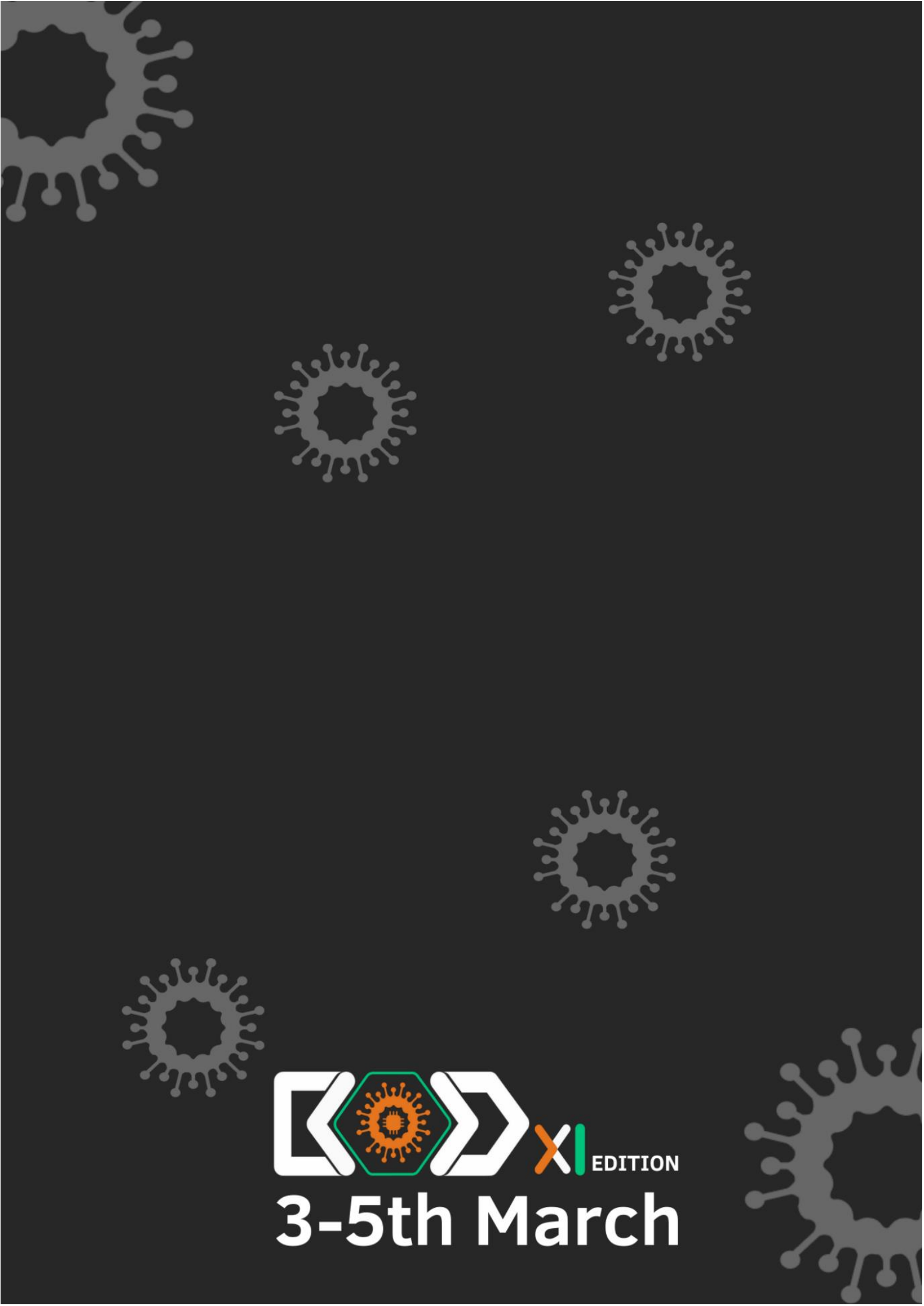
Furthermore, in the development of human understandable explanations, one of the main factors to consider is the user experience and human-understanding of the explanations[8]. One aspect that improves this experience is the ability to visualise the explanations in a clear and comprehensible way[9].

Taking the context and problems discussed above, the goal of this work is to create a visualisation tool to explain AI clinical predictions from black-box models. Using a knowledge graph to map the available data, the goal is to provide the best explanation possible for each prediction and to represent it with a visual tool developed for this purpose. This is a novel approach since it combines semantic explanations with visualisation. There have been several different approaches in the field of explainable AI, but they are mainly divided into either the use of knowledge bases, or visual explanations. This work will combine both approaches to hopefully obtain better explanations.

Acknowledgements The KATY project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101017453. The content provided in this Abstract reflects the author's views only. The work is also funded by the FCT through LASIGE Research Unit, ref. UIDB/00408/2020 and ref. UIDP/00408/2020.

References

1. Min, S., Lee, B. & Yoon, S. Deep learning in bioinformatics. *Brief. Bioinform*. 18, 851–869 (2017).
2. Clinciu, M. A. & Hastie, H. F. A survey of explainable AI terminology. NL4XAI 2019 *-1st Work. Interact. Nat. Lang. Technol. Explain. Artif. Intell. Proc. Work*. 8–13 (2019)doi:10.18653/v1/w19-8403.
3. Holzinger, A., Biemann, C., Pattichis, C. S. & Kell, D. B. What do we need to buildexplainable AI systems for the medical domain? 1–28 (2017).
4. Guidotti, R. *et al*. A survey of methods for explaining black box models. ACM *Comput. Surv*. 51, (2018).
5. GPDR.eu. General Data Protection Regulation (GDPR) Compliance Guidelines. *gpdr.eu* https://gdpr.eu/ (2020).
6. Chari, S., Gruen, D. M., Seneviratne, O. & Mcguinness, D. L. Foundations of Explainable Knowledge-Enabled Systems. (2020).
7. Leuce F. On the Role of Knowledge Graphs in Explainable AI. *Semant. Web* 1, 1–5 (2020).
8. Wang, D., Yang, Q., Abdul, A., Lim, B. Y. & States, U. Designing Theory-Driven User-Centric Explainable AI. 1–15 (2019).
9. Öztürk, Ö. & Açikgöz, H. G. Onyx: A new Canvas-based tool for visualising biomedical and health ontologies. *undefined* 37, (2020).


XI EDITION
3-5th March