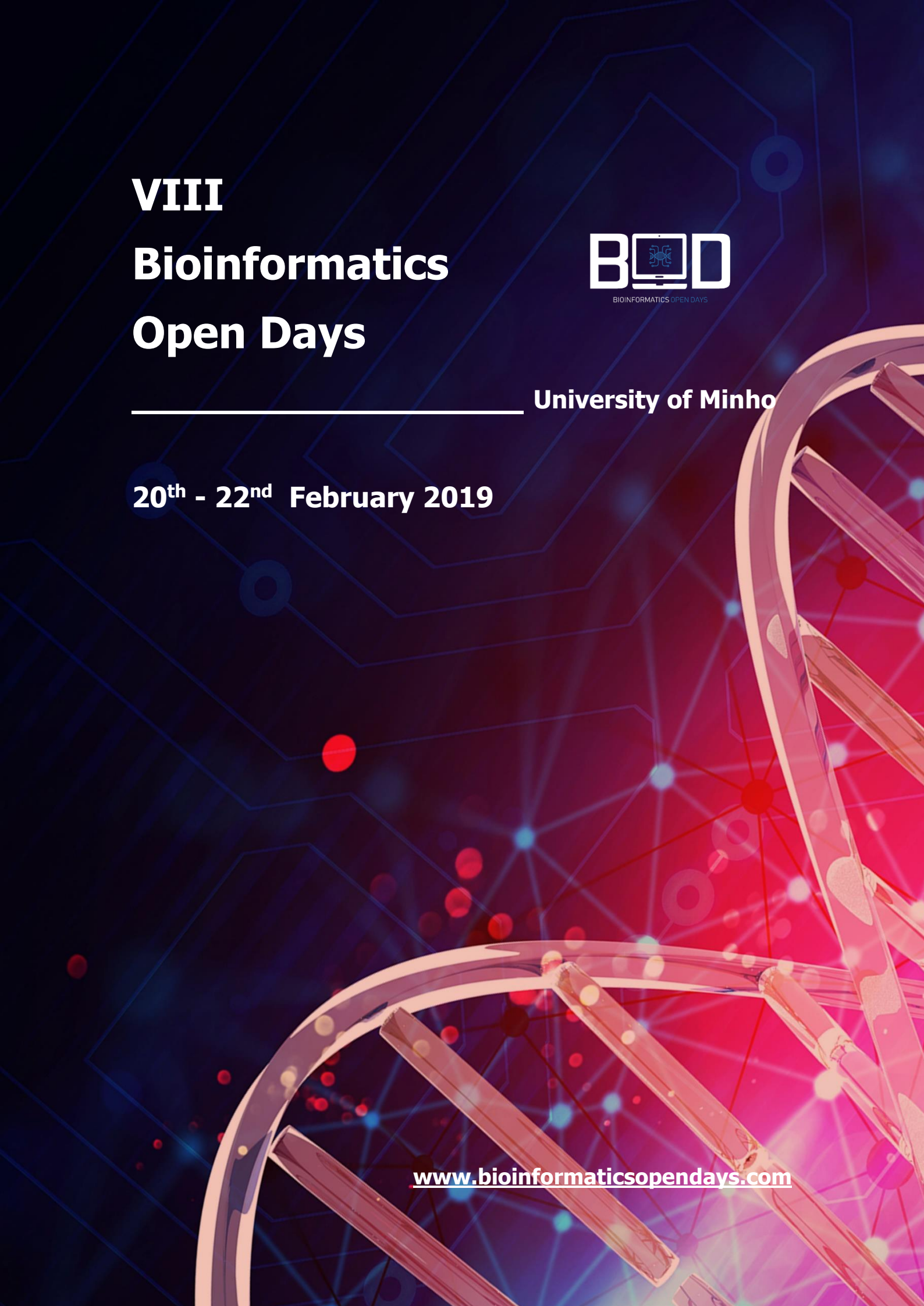
VIII Bioinformatics Open Days

BIOINFORMATICS OPEN DAYS

University of Minho

20th - 22nd February 2019



www.bioinformaticsopendays.com

21st February

Lecture 1: SENSING ADAPTATION WITH GENOMICS AND BIOINFORMATICS

Agostinho Antunes

CIIMAR, Interdisciplinary Centre of Marine and Environmental Research, University of Porto, Portugal

Department of Biology, Faculty of Sciences, University of Porto, Portugal

Whole genome sequencing projects can be crucial to decipher genetic disease and health, species evolution and the diversification of phenotypic traits. Here, recent results from our group retrieved from comparative evolutionary genomic analyses of varied animal species will exemplify adaptive successes to thrive into diverse ecological environments. The findings pinpoint unique molecular products of critical relevance in species evolution, diversification and conservation, but also highlight genomic novelties with relevance in environmental and biomedical research.

Oral Presentations

1. EvoPPI 1.0: a Web Platform for Within and Between Species Multiple Interactome Comparisons and Application to Nine polyQ Proteins Determining Neurodegenerative Diseases

Cristina P. Vieira[1,2], Sara Rocha[1,2], Noé Vázquez[3,4], Hugo López-Fernández[1,2,3,4,5], André Torres[1,2], Rui Camacho[6], Florentino Fdez-Riverola[3,4,5], Miguel Reboiro-Jato[3,4,5], Jorge Vieira[1,2]

1 - Instituto de Biologia Molecular e Celular (IBMC), Rua Alfredo Allen, 208, 4200-135 Porto, Portugal

2 - Instituto de Investigação e Inovação em Saúde (I3S), Universidade do Porto, Rua Alfredo Allen, 208, 4200-135 Porto, Portugal

3 - ESEI - Escuela Superior de Ingeniería Informática, Edificio Politécnico, Campus Universitario As Lagoas s/n, Universidad de Vigo, 32004 Ourense, Spain

4 - Centro de Investigaciones Biomédicas (Centro Singular de Investigación de Galicia), Vigo, Spain
5 - SING Research Group, Galicia Sur Health Research Institute (IIS Galicia Sur). SERGAS-UVIGO 6LIAAD & DEI & Faculdade de Engenharia, Universidade do Porto, Portugal

Protein-protein interaction (PPI) data is essential to elucidate the complex molecular relationships in living systems, and thus understand the biological functions at cellular and systems levels. The complete map of PPIs that can occur in a living organism is called the interactome. For animals, PPI data is stored in multiple databases (e.g. BioGRID, CCSB, DroID, FlyBase, HIPPIE, HitPredict, HomoMINT, INstruct, Interactome3D, mentha, MINT, and PINA2) with different formats. This makes PPI comparisons difficult to perform, especially between species, since orthologous proteins may have different names. Moreover, there is only a partial overlap between databases, even when considering a single species. The EvoPPI (http://evoppi.i3s.up.pt) web application here presented allows to compare the data from the different databases at species level, or between species by using a Blast approach. We show its usefulness by performing a comparative study of the interactome of the nine polyglutamine (polyQ) disease proteins namely androgen receptor (AR), Atrophin-1 (ATN1), ataxin 1 (ATXN1), ataxin 2 (ATXN2), ataxin 3 (ATXN3), ataxin 7 (ATXN7), calcium voltage-gated channel subunit alpha1 A (CACNA1A), Huntingtin (HTT), and TATA-binding protein (TBP). Here we show that none of the human interactors of these proteins is common to all nine interactomes. Only 15 proteins are in common with at least four of these polyQ disease proteins, and 40% of these are involved in ubiquitin protein ligase binding function. Although most of the polyQ disease proteins are transcription factors, the results here obtained suggest that they are involved in different functional networks. Comparisons with Mus musculus PPIs are also made for AR and TBP, using EvoPPI BLAST search approach (a unique feature of EvoPPI), with the goal of understanding why there is a significantly large number of common interactors observed between these proteins in humans. To address the effect of the polyQ region in such pattern, the mean number of interaction in polyQ proteins and non-polyQ proteins have been determined in H. sapiens, M. musculus, R. norvegicus, B. taurus, D. melanogaster, and C. elegans (all animal species having more than 10000 interactions in at least one interactome database). A clear effect of the polyQ on the number of interactors is observed in humans only, and thus, we

cannot exclude the effect of the polyQ region on the large number of common interactors between AR and TBP.

2. Implementation of Genomic Selection in the Portuguese wheat breeding program

Octávio Serra[1], Sílvia C. Alves[1], Fernanda Simões[1,2], Ana Sofia Almeida[3], Ana Rita Costa[3], Nuno Pinheiro[3], José Coutinho[3], Conceição Gomes[3], Nuno Carolino[1], Inês Carolino[1], Daniel Gaspar[4], Carla Borges[3], José Moreira[3], Paula Scotti[3], José Semedo[3], Isabel Pais[1,2], Marcos Ramos[4], José Matos[1,2], Benvindo Maçãs [3]

1 - Instituto Nacional de Investigação Agrária e Veterinária, Quinta da Fonte boa, 2005-048 Vale de Santarém;

2 - Instituto Nacional de Investigação Agrária e Veterinária, Avenida da República, Quinta do Marquês, 2784-505 Oeiras;

3 - Instituto Nacional de Investigação Agrária e Veterinária, Estrava de Gil Vaz, Apartado 6, 7351-901 Elvas;

4 - CEBAL, Rua de Pedro Soares, 7800-309 Beja

In spite of climate change, due to the estimated increase in Human population in the next decades, farmers are expected to proportionally deliver additional high-quality food at low prices, while employing fewer natural resources, and protecting the bio-ecosystem from soil erosion, chemical fertilizers, and pesticides.

Among the world's top 10 staple foods, wheat is the third main crop cultivated worldwide; and its global demand is increasing, as a result of the worldwide industrialization process and diet westernization.

In this context, breeding outperforming varieties in a changing climate, while demanding fewer inputs, is a strategic imperative.

The Portuguese Estação de Melhoramento de Plantas (INIAV, Elvas) has been the national plant breeding center since 1942. Today, however, the conventional breeding techniques are no longer adequate for modern-day demands. These outdated trials are very expensive and take, at least, over one decade to eventually produce an improved variety.

Genomic Selection is a new approach to improve quantitative traits that combines genome-wide marker data with phenotypic and pedigree data to infer the genetic potential of each individual parent, and predict which crosses will produce the highest increase in genetic gain in the shortest amount of time.

This tool relies on Genotyping By Sequencing to produce, via next generation sequencing, a high number of molecular markers in two types of datasets: a training set and a validation set. The breeding values of new genotypes are estimated based on the marker effects predicted by a statistical model fitted to the training set, and independently supported by the validation set - which contains genotyped individuals, derived from the reference population, that haven't been phenotyped.

The FASTBREED project (ALT20-03-0145-FEDER-000018) is pioneer in Portugal by introducing this bioinformatics approach into the national wheat breeding program. This represents a major turning point for the Portuguese wheat farmers in need of newly improved and well-adapted varieties to face the challenges brought by climate-change.

3. Construction of a No-SQL database for variant allele frequencies of 70 Portuguese samples

Daniel Martins[1,2]; Hugo Froufe[1]; Diogo Pinho[1]; Cristina Barroso[1]; Maria José Simões[1]; Conceição Egas[1,3].

1 – Genoinseq - Next Generation Sequencing Unit, Biocant, BiocantPark, Núcleo 04, Lote 8, 3060-197 Cantanhede, Portugal;

2 - Universidade do Minho, Largo do Paço, 4704-553 Braga, Portugal;

3 - Center for Neuroscience and Cell Biology, University of Coimbra, 3004-504 Coimbra, Portugal.

As NGS technology generates enormous amounts of data, efficient infrastructures to store and retrieve that information become essential for managing and interpret genetic data. New genetic information is particularly relevant for research and clinical purposes.

Global-scale initiatives provide a reference for genomic studies. However, to our knowledge, the two most complete populational projects, the 1000 Genomes Project (1kG) and gnomAD, do not include any Portuguese sample. We believe that a Portuguese

collection of genomic information would greatly benefit molecular diagnosis in Portuguese patients.

We constructed a No-SQL Database with more than 200,000 variants detected in 70 Portuguese individuals inserted along populational information and genomic annotation.

Each variant was normalized, thus being represented by the minimum number of nucleotides required to present the alteration. For each variant, the genotypes called for all covered samples were registered, allele frequency and sample and genotype counts were calculated and inserted on the database.

All processed variants were stored independently and searched on both 1kG and gnomAD. Information for corresponding variants on both sources was added to the database. Variants were annotated with effect predictions, gene name and clinical outcomes.

The database enabled populational comparisons to various populations, which endorsedthe evidence for a correlation between genetic and geographic distance previously reported in literature.

Significative differences for allele distribution were found between our population and the other 1kG European subpopulations in 7,284 variants distributed by 2,571 genes. Results suggest the existence of populational genetic markers and may prompt future studies for detection of Portuguese-specific genetic markers.

This work was supported by the In2Genome project CENTRO-01-0247-FEDER-017800, supported by Centro Portugal Regional Operational Programme (CENTRO 2020), under the Portugal 2020 Partnership Agreement, through the European Regional Development Fund (ERDF).

4. HIV-1 subtyping by phylogenetic pairing: a snakemake pipeline

Pedro M.M. Araújo[1,2], Nuno S. Osório[1,2]

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.

The main obstacles for the Human immunodeficiency virus 1 (HIV-1) eradication are the extensive viral genetic diversity and fast adaptive capacity, consequence of high viral replication, recombination, and mutation rates. This vast diversity led to the

necessity of classifying HIV-1 in several groups, subtypes and sub-subtypes. The clinical relevance of HIV-1 subtypes is still not fully understood. However, studies suggest that distinct subtypes may lead to differences in the disease progression rate, advantages in specific transmission routes, different predispositions to evade the immune response, and even differences in therapeutic outcomes. Therefore, HIV-1 subtyping is of major importance for surveillance studies to better understand and control the global epidemic. HIV-1 sequencing (partial genome) is routinely done in healthcare facilities to monitor drug resistance polymorphisms, creating large datasets for subtyping. According to recent reports the HIV-1 subtype diversity is increasing in

Europe and other regions, making its motorization even more relevant. HIV-1 subtyping is mainly done using web servers hosted in different regions of the globe. These tools have limitations in scalability and might raise problems in compliance with data protection legislation. Therefore, there is a need for an HIV-1 subtyping solution which is available to be run locally, is scalable, efficient, and reliable. To answer this we designed a Snakemake-based pipeline for HIV-1 subtyping that works by phylogenetic paring. Snakemake creates a directed acyclic graph (DAG) of jobs allowing the efficient parallelization of tasks. After mapping to the reference genome the given target is added to a previously created phylogenetic tree backbone, then the closest phylogenetic pair, the node most recent ancestors, and the branch support values are inferred. This information is then made available to the user in order to make the subtyping decision as informative as possible.

BioData.pt Special Session

1. BioData.pt/ELIXIR Portugal: A Biological Data Infrastructure for Research and Innovation

Mário Gaspar da Silva[@1, 2] and Ana Portugal Melo[#1, 3]

@ - Head of Node of ELIXIR Portugal and Vice-President of BioData.pt

- Deputy Head of Node of ELIXIR Portugal and Executive Director of BioData.pt

1 – BioData.pt

2 – INESC-ID and IST, U. Lisboa

3 – Portuguese Infrastructure of Biological Data – BioData.pt

BioData.pt is the distributed Portuguese National Infrastructure of Biological Data for the scientific and industrial biotech community. Biodata.pt is a virtual organization, dedicated to serving scientific and business research and innovation, pooling the resources of 12 institutions, from North to South of Portugal, dedicated to leveraging value creation initiatives based on biological information. The consortium promotes interchanges with the Academic and Business clusters, making Research accessible to Innovation, namely in the agri-food and forestry, sea and health sectors. Complementing its national dimension, BioData.pt is the Portuguese Node of ELIXIR, the European Infrastructure for the Life Sciences, a European intergovernmental organization that brings together researchers in the areas of life sciences and computing, with the aim of helping researchers academics and businesses to take advantage of the large amounts of data produced by biological research

2. The training platform of BioData.pt

Pedro L. Fernandes

Portuguese Infrastructure of Biological Data - BioData.pt, Instituto Gulbenkian de Ciência, Oeiras, Portugal

The training platform of BioData.pt is acting in three main trends:

1. Organising documentation about courses in the Gulbenkian Training Program in Bioinformatics (GTPB), to allow for the replication of the training model;
2. Enhancing of the Bioinformatics Training Room at the IGC to enable more interaction, better audiovisual means, higher comfort for course participants.
3. Attracting ELIXIR training activities to Portugal, both in the scope of the Use Cases in ELIXIR and the prioritised areas of the Training Workpackage.

New themes for training are settling-in such as Computational Pangenomics, Integrative Proteomics and Data Management & Stewardship. Furthermore, some of the training will specifically target industrial users of the BioData.pt infrastructure, meeting their interests in a more flexible format. We are also stepping into distance and e-learning, authoring materials for inclusion in Learning Management Systems.

3. BioData.pt Computing Services & Infrastructure

Luís Guerra e Silva

Assistant Professor of the Department of Computer Science and Engineering of IST / University of Lisbon. VP of IST or Information and Communication Technologies. Senior researcher at the Algorithms for Optimization and Simulation Group (ALGOS) of INESC-ID.

In this talk we will present a brief overview of existing BioData.pt computing services and will discuss the underlying infrastructure. We will also present new services, to be made available during 2019, which are currently under development.

4. Biodata.pt/ELIXIR Portugal: A Data Management Service for Biodata.pt

João Cardoso[@#], Alexandre Francisco[@#], José Borbinha[@#]

@ - INESC-ID, Instituto Superior Técnico, Universidade de Lisboa, Lisboa, Portugal

- Portuguese Infrastructure of Biological Data – BioData.pt, Computing Platform

The objective of BioData.pt is to provide to the Portuguese scientific community and industry with a distributed e-infrastructure of services for biological data management processing. BioData.pt is also the Portuguese node of ELIXIR. The services for data management will comprise data storage, sharing and data life-cycle management in general, according to a Data Management Plan (DMP). A DMP must define the requirements for the data life-cycle management, such as administrative data, data access, data sharing, data preservation, policies, etc. The research challenge in this scenario will rest on how individual users or user communities create a DMP. Either by starting from scratch or by reusing any existing DMP documents (BioData.pt will therefore also provide and manage references for a DMP document). BioData.pt will support the registration of external data processing services, which can then be reused in processing workflows (According to the terms of services. E.g. if they are of open public use, require previous licensing, etc.). BioData.pt also will allow recording, management, reuse and sharing of these workflows, thus making them reproducible. Concerning the technology, BioData.pt will comprise a digital repository supported by Invenio, the open source software framework originally developed for the CERN Open Data service, and now used for many other services, such as Zenodo . The data processing services will make it possible to create virtual machines in OpenStack-based Cloud Services. In this sense, BioData.pt will support the Portuguese scientific community to follow the FAIR data

principles , while also accommodating other scenarios if required by reasonable industry's business models. We also expect BioData.pt to also be a reference for the new RDA-Portugal working group6 , as either a motivation for thinking, discussion and learning, and as a platform available for experiments and demonstrations concerning scientific data management. This presentation will describe in more detail the vision for the BioData.pt services, the current status, and planned future steps. It also will be motivated feedback from the audience of the session.

5. One year of bioinformatics user-support within BioData.pt

Daniel Pinto-Neves[1,2], Ricardo Leite[1], Daniel Sobral[1,2]

1 - Instituto Gulbenkian de Ciência, Oeiras, Portugal

2 - Portuguese Infrastructure of Biological Data – BioData.pt

BioData.pt is the Portuguese node of ELIXIR, the European intergovernmental infrastructure for biological data. As part of this infrastructure, the BioData.pt bioinformatics support workgroup, operating from two nodes, one in Instituto Gulbenkian de Ciência (IGC), Oeiras, and the other in Instituto de Biologia Molecular e Celular, Porto, aims at empowering researchers in the lifesciences, both in national academic institutions and in industry, with training, consulting services, ad-hoc analysis of biological data, as well as access to computing infrastructure and tools. Here we present our four-pronged approach to bioinformatics support, that includes 1. direct consulting and data analysis services; 2. production and delivery of training materials (in coordination with the training workgroup); 3. development of user-friendly and accessible user interfaces for common data analysis tools; and 4. setup and deployment of value-added computational resources, such as virtual machines pre-configured with common analysis pipelines, through a cloud infrastructure. Throughout our first year we have provided direct support to research groups in several Portuguese institutions, including Instituto de Medicina Molecular, Fundação Champalimaud and Universidade do Porto. We have delivered training courses and workshops (in conjunction with the GTPB program at the IGC and The Carpentries) in topics such as RNA-seq and single-cell RNA-seq data analysis, data processing and genomics. We continuously strive to incentivize user autonomy, and for this we have started development of two web applications (based on the R Shiny framework) to make common bioinformatics analyses more accessible to non-computer scientists.

6. Cork oak genomics in the post-genome sequence era

A.M. Ramos[1,2,3] Genosuber Consortium[4,5,6,7]

1 - CEBAL - Centro de Biotecnologia Agrícola e Agro-Alimentar do Alentejo, Rua Pedro Soares, 7801- 908 Beja, Portugal

2 - Instituto de Ciências Agrárias e Ambientais Mediterrânicas (ICAAM), Universidade de Évora, Évora, Portugal

3 - Portuguese Infrastructure of Biological Data – BioData.pt

4 - ITQB - Av. da República, Estação Agronómica Nacional, 2780-157 Oeiras, Portugal

5 - IBET - Av. República, Qta. do Marquês, Estação Agronómica Nacional, Edificio IBET/ITQB, 2780- 157 Oeiras, Portugal

6 - INIAV - Av. da República, Quinta do Marquês, 2780-157 Oeiras, Portugal

7 - Biocant - Parque Tecnológico de Cantanhede, Núcleo 04, Lote 3, 3060-197 Cantanhede, Portugal

The Genosuber project, a consortium formed by several Portuguese institutions, has determined the genome sequence of cork oak, one of the main forestry species in the Mediterranean, which plays a relevant biological and economic role in Portugal. The outstanding advances observed in high-throughput sequencing allow the generation of substantial volumes of sequence data. However, without a sequenced reference genome these powerful datasets cannot be fully explored. With the availability of a fully sequenced and annotated genome, cork oak genomics research can now enter a new era.

In addition to the cork oak genome assembly, several omics datasets focusing on different types of sequencing have been produced in Genosuber, with the purpose of advancing cork oak genomics research applied to relevant biological systems and phenotypic traits. Large scale SNP detection projects are underway, based on whole-genome resequencing of cork oak trees, with contrasting phenotypes for cork quality, and hundreds of thousands of SNPs were identified. These SNPs are being used to perform genome-wide association studies. Moreover, structural variation and copy number variation have also been identified with these WGRS datasets. Furthermore, the selective sweeps that were identified provide the first insight regarding cork oak evolutionary and selection processes.

Cork oak transcriptomics was previously studied, focusing on the identification of differentially expressed genes and functional characterization of transcriptomes. More

recently, studies targeting the cork oak non-coding transcriptome have also been executed, including the analysis of micro and long non-coding RNAs. Within Genosuber, a comprehensive transcriptomic characterization of tissues involved in cork formation was produced. The epigenomic landscape of these tissues was also determined, which added additional layers of omics information, with great potential to unravel the regulation of this complex trait.

These studies have clearly indicated that it will be possible to apply all types of omics approaches to cork oak research. However, they have also demonstrated the need for a database resource where all this information can be stored, curated and maintained. This resource, being developed within the Biodata project, will be crucial to agglomerate all cork oak omics results and accelerate future research studies in the species.

7. The Cork Oak Genome Database

Cirenia Arias-Baldrich[1,2], Célia Miguel[2,3,4], Marcos Ramos[5], Daniel Sobral[1,2]

1 - Instituto Gulbenkian de Ciência, Oeiras, Portugal

2 - Portuguese Infrastructure of Biological Data – BioData.pt

3 - Instituto de Biologia Experimental e Tecnológica, Oeiras, Portugal

4 - Biosystems & Integrative Sciences Institute, Faculdade de Ciências da Universidade de Lisboa, Lisboa, Portugal

5 - Centro de Biotecnologia Agrícola e Agro-Alimentar do Alentejo, Beja, Portugal

Quercus suber L., commonly known as cork oak, is an evergreen tree species native to Southwest Europe and North Africa. This species has a key role in the ecology and economy of these areas, among other reasons, on account of cork production. Deciphering the mechanisms of Q. suber's response to the environment and getting a deep insight into its biology is crucial to counteract abiotic and biotic stresses that affect this tree and can compromise the stability of its ecosystem. A first draft version of the cork oak genome has been released recently and will be key in helping researchers to address this challenge. In an effort to integrate this information in a comprehensive, accessible and intuitive format, we are developing the Cork Oak Genome Database web portal. This portal is based on Tripal, which is an open-source toolkit for construction of online biological and genomic databases. The Cork Oak Genome Database is supported by BioData.pt, the Portuguese distributed e-infrastructure for biological data, and Portuguese node of ELIXIR. BioData.pt bridges academy and industry and its efforts include making

research available for innovation. The Cork Oak Genome Database will provide a service for users to search and explore the available genomic and transcriptomic data on cork oak. Some functionalities such as search, a JBrowse link to visualize RNAseq data available for five cork oak tissues (leaf, pollen, phellem, xylem and innerbark), and straight-forward access to InterProScan and Blast Analysis results, are already implemented. We are currently working in providing a userfriendly interface, publicly available and with functional tools to help the research community take advantage of growing genomic information available.

8. MIAPPE 1.1: Towards a New Generation of Metadata Standards

Daniel Faria[1] , Inês Chaves[2,3], Bruno Costa[3,2], Célia Miguel[2,3,4]

1 - Portuguese Infrastructure of Biological Data – BioData.pt, Instituto Gulbenkian de Ciência, Oeiras, Portugal

2 - Instituto de Biologia Experimental e Tecnológica, Oeiras, Portugal

3 - ITQB NOVA, Oeiras, Portugal

4 - Biosystems & Integrative Sciences Institute, Faculdade de Ciências da Universidade de Lisboa, Lisboa, Portugal

Metadata standards are a key component of the FAIR data principles. These standards detail the information that must be provided about a dataset—which helps address both findability and reusability—and also detail how that information should be organized and represented—which helps address interoperability.

The need for metadata standards in the life sciences was first recognized with the advent of DNA microarrays, and led to the creation of the MIAME (Minimum Information About a Microarray Experiment) standard and its data exchange format MAGE-tab. Several other metadata standards have since been developed for the life sciences, and MAGE-tab has given rise to the more generic ISA-tab, which can be used with a number of different standards.

In the past three years, the European life science network, ELIXIR, and its Portuguese representative, BioData.pt, undertook the task of extending the metadata standard for plant phenotyping experiments, MIAPPE (Minimum Information About a Plant Phenotyping Experiment). This extension, which was recently accepted by the community as MIAPPE 1.1, included a broadening of scope to accommodate woody plants as an additional use-case, as well as the enrichment of field definitions and

examples to improve accessibility. More importantly, it included the specification of a data model, to improve usability and enable automatic validation, and mappings between this data model and both ISA-tab and BrAPI, to ensure that both can be used interoperably to record MIAPPE datasets. Last, but not least, it included the encoding of the MIAPPE standard and data model in OWL, to support the expression of MIAPPE datasets in RDF, and thus bring metadata standards to the age of linked data.

9. The Metabolic Model of the *Quercus suber* (Cork Oak Tree) Leaf

Hüseyin DEMIRCI[1], Oscar DIAS[1], Inês CHAVES[2], Célia Miguel[2], Miguel ROCHA[3], Isabel ROCHA[2]

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

2 - ITQB NOVA, Av. da República, 2780-157 Oeiras, Portugal

3 - Department of Informatics, University of Minho, 4710-057, Braga, Portugal

The cork oak tree, Quercus suber, is an important renewable resource from which wine stoppers and many other natural products are derived. Portugal is the main producer of cork which approximately manifactures half of the world's total consumption. The recently sequenced genome of Quercus suber [1] has 953 Mb size containing about 79.000 genes. It is important to have a better under- standing of the genomics and metabolomics of the tree to increase resistance to abiotic and biotic stresses and to obtain high quality cork. This information can be used as a supporting parameter for decision-making in cork production since normally it is not possible to evaluate the quality of cork before the tree is 40 years old.

In this work we present a metabolic genome scale model for the leaf of the Quercus suber. We have used the Merlin software [2] to reconstruct the draft metabolic model. The enzymes are annoted using UniProtKB and SwissProt databases [3]. For the annotation we have used previously annotated plant en- zymes from species such as Quercus, Arabidopsis thaliana, Oryza sativa subsp. japonica, Vitis vinifera, Zea mays and Solanum tuberosum using the Blast e- value threshold e−10. If there was no suitable plant species, we annotated the enzyme with any other organism with e-value < e−10. As it is a well studied model organism, the majority of the enzymes have been annoted with Ara-bidopsis thaliana. More than 1600 annotations were manually curated with the help of Merlin's environment. The major pathways including Glycolysis/ Gluconeogenesis,

carbon fixation, TCA cycle, and production of amino-acids and other biomass precursors have been checked manually. The biochemical reactions were checked to be chemically and electrically balanced. Mostly, H+ and H2O may be missing in some databases therefore the reactions were updated according to MetaNetX, KEGG, BiGG databases when related information is found. Also the gaps in the pathways have been investigated and removed by adding/correcting necessary equations.

The biomass composition has been defined using protein, carbohydrates, lipids, cofactors, DNA and RNA components using similar approaches in previous plant models such as AraGEM and Tomato [4, 5]. The required drains for photons and inorganic compounds for CO2,H2O,O2 are defined in the model to describe the uptake/secretion of these compounds. The sources for Nitrogen, Phosphate, Sulphur are also defined similarly with the help of Ammonia, Nitrate, Orthophosphate, Sulphate and Hydrogen sulfide.

The obtained Cork model consists of 3269 reactions, 2934 metabolites and 7531 genes of which 405 are transporters. The transport genes have been identi- fied using the Triage tool of Merlin. The compartment prediction has been using Loc3Tree [6] protein localization prediction system. Chloroplast, cytoplasm, endoplasmic reticulum, golgi apparatus, mitochondrion, nucleus, peroxisome, plasma membrane, plastid, and vacuole are the predicted locations inside the plant cell. For the reversibility of the directions, we forced the Kegg's reaction directions according to the the yeast model reactions. Later we have manually curated the directions to guarantee the growth of biomass precursors.

The validity of the model has been checked for biomass production, using simulation tools such as Optflux [9]. Also, inhouse model validity tools have been used for the control of biomass precursors. The derived Cork model is able to grow biomass and produce Oxygen under photosynthetic conditions with the help of photons where CO2 is the main carbon source. On the other hand when the main carbon source is defined as glucose, the model can simulate the respiration, producing CO2 and H2O and consuming Oxygen. We observe non-zero flux at about 1445 reactions which is also a comparable result with previous studies. We conclude that our leaf model is capable of simulating both photosynthetic (light) and respiration (dark) reactions. To the best of our knowledge, this is the first metabolic model of a tree species. We believe that this model will bring new insights for Quercus suber studies such as the formation process of cork. Also, it will be a basis for metabolic models of other plant species. Considering plants as resources of many valuable natural metabolites, these models can bring significant biotechnological applications.

References

[1] The draft genome sequence of cork oak, Ramos Antonio Marcos et al., Scientific Data, 5, 2018.

[2] Reconstructing genome-scale metabolic models with merlin,Oscar Dias, Miguel Rocha, Eug´enio C. Ferreira, Isabel Rocha; Nucleic Acids Research, Volume 43, Issue 8, pp. 3899–3910, 2015.

[3] The UniProt Consortium, UniProt: the universal protein knowledgebase , Nucleic Acids Res. 46: 2699 (2018). [4] AraGEM, a Genome-Scale Reconstruction of the Primary Metabolic Network in Arabidopsis, Cristiana Gomes de Oliveira Dal'Molin, et al., Plant Physiology, 152 (2):579-589, 2010.

[5] A genome-scale metabolic network reconstruction of tomato (Solanum ly- copersicum L.) and its application to photorespiratory metabolism, Yuan H. et al., Plant J., 85(2):289-304, 2016.

[6] Loc3Tree: Protein Subcellular Location Prediction Sytem, Goldberg T. et al. Nucleic Acid Research, 2014.

[7] Responses to lightintensity in a genome-scale model of rice metabolism, Poolman, M.G. et al., Plant Physiol. 162, 1060–1072, 2013.

[8] Assessing the metabolic impact of nitrogen availability using a compart- mentalized maize leaf genome-scale model, Simons M. et al., Plant Phys- iol., 166(3):1659-74, 2014.

[9] OptFlux: an open-source software platform for in silico metabolic engi- neering, Isabel Rocha et al., BMC Systems Biology 4:45, 2010.

Panel Discussion: Learning Bioinformatics in Portugal

Miguel Rocha[1], Irene Oliveira[2], Octávio Paulo[3]

1 - Director of the master in Bioinformatics at University of minho

2 - Director of the master in Bioinformatics at University of Trás-dos-Montes e Alto Douro

3 - Coordinator of the master in Bioinformatics and Computational Biology at University of Lisbon

Informal conversation with the participation of Bioinformatics Professors from different national academies, with the purpose of assessing the current picture of Bioinformatics in Portugal, mainly its evolution through time, the ongoing state of education in this area, as well as other relevant subjects.

22nd February

Lecture 2: Genomic Stories of DNA Mutation and Damage

Nicholas Luscombe

Francis Crick Institute

University College London Genetics Institute

Okinawa Institute of Science & Technology

Nicholas Luscombe has a degree in natural sciences from the University of Cambridge, a PhD on the basis for specificity of DNA-binding proteins from the University College London, and a postdoctoral fellow on yeast transcriptional regulation from the University of Yale. In 2005 he became group leader at the EMBL-European Bioinformatics Institute where he stayed until 2012 and built a computational biology laboratory with an emphasis on genomics and gene regulation. During this time, he joined the Okinawa Institute of Science & Technology as an Adjunct Faculty to establish a small group focused on developmental regulation (2011-present). Currently, he teaches Computational Biology in the University College London Genetics Institute and holds the position of Senior Winton Group Leader at the Francis Crick Institute in London.

Oral Presentations

1. Structural Bioinformatics studies on the Influenza virus fusion machinery

Diana Lousa[1], Antónia R. T. Pinto[2], Bruno L. Victor[1], Alessandro Laio[3], Sara R. R. Campos[1], António M. Baptista[1], Ana S. Veiga[2], Miguel A. R. B. Castanho[2] and Cláudio M. Soares[1]

1 - ITQB NOVA, Instituto de Tecnologia Química e Biológica António Xavier, Universidade Nova de Lisboa, Av. da República, 2780-157 Oeiras, Portugal

2 - Instituto de Medicina Molecular, Faculdade de Medicina de Lisboa, 1649-028 Lisboa, Portugal

3 - SISSA/ISAS, Statistical and biological physics, Via Beirut 2-4 Trieste, Italy

Influenza pandemics are some of the most serious health threats of our time, in view of the limited treatments available. Research on the molecular mechanisms of infection by the influenza virus (IV) is needed to find new therapeutic targets. Inactivating the fusion of the viral and host membranes is considered a promising strategy, but this process is poorly understood at the molecular level.

The IV fusion process is promoted by the protein hemagglutinin (HA). The IV is uptaken by endocytosis and the low pH of the late endosome triggers a large conformational change of HA that initiates fusion. HA contains a key regions that is essential for this process: the fusion peptide (FP). The FP binds to the host membrane and promotes fusion. Interestingly, this peptide is able to induce fusion of lipid vesicles, even in the absence of the rest of the protein, making it a privileged model to study fusion.

In the last years, our group has studied the molecular determinants of the FP activity. Using state of the art simulation methods, we have shown that this peptide can adopt two different conformations in the membrane, which have different impacts on the membrane properties. Our work also shed light into the mechanisms by which the peptide perturbs the membrane. To understand the role played by key residues, we have performed bias-exchange metadynamics simulations of different FP mutants, which allowed us to characterize their energy landscape and provide insights into the effect of mutations on the peptide activity. Another important question that we are addressing concerns the effect of pH on the peptide's structure and membrane-interacting properties. By combining the simulation results with experimental studies performed by our collaborators, we were able to provide a detailed molecular characterization of the influenza FP, which can be useful for the design of novel therapies against this devastating pathogen.

2. A practical comparison of analytical strategies for microbial bioinformatics data analysis

Vera Manageiro[1], Miguel P. Machado[2], Manuela Caniça[1], João A. Carriço[2]

1 - National Reference Laboratory of Antibiotic Resistances and Healthcare Associated Infections (NRL-AMR/IACS), Department of Infectious Diseases, National Institute of Health Dr Ricardo Jorge (NIH), Lisbon, Portugal

2 - Instituto de Microbiologia, Instituto de Medicina Molecular, Faculdade de Medicina, Universidade de Lisboa, Lisbon, Portugal

Whole-genome sequencing (WGS) is highly discriminative for molecular typing of bacterial isolates and offers an unprecedented resolution for tracking outbreaks and performing surveillance at national or global scales. The aim of this study was to explore the utility of WGS for Enterobacteriaceae outbreak investigations by comparing two analytical strategies for bioinformatics data analysis.

WGS of 33 Klebsiella pneumoniae isolates belonging to the strain collection of the NRL-AMR/IACS, from NIH, Lisbon, was performed on Illumina MiSeq. In the first approach, we used the bioinformatics pipeline INNUca for performing quality control of reads, de novo assembly and contigs quality assessment. Prokka and ABRicate software were used for genome annotation and screening for antimicrobial resistance and/or virulence genes, respectively. Phylogenetic inference was performed with chewBBACA pipeline. Two-hundred thirty-six NCBI publicly available K. pneumoniae genomes used to define the whole/core genome MultiLocus Sequence Typing (wg/cgMLST) schema. The evolutionary relationship was accessed by Minimum spanning trees (MST) using the PHYLOViZ. In the second approach, raw reads were analyzed with the BioNumerics wgMLST pipeline (Applied Maths). Assembly-free allele detection analyses were performed for each isolate. The assembled genomes were analyzed using the E. coli genotyping plug-in (serotype, virulence, and resistance prediction). Phylogeny was also inferred by calculating MST.

The results of the analysis which included prediction of antimicrobial resistance phenotype and virulence gene detection, Klebsiella locus typing, seven-gene MLST, cgMLST and phylogenetic inference, showed an overall good correlation between the data produced using the open source pipeline and those with BioNumerics. Both cgMLST approaches yield a high discriminatory power for comparison of isolates when compared the traditional seven-gene MLST for the characterization and typing of isolates. However, allelic profiles generated were different according to the scheme used by each method. The commercial platform BioNumerics had the advantage of a single user-friendly platform, after a certain level of training. The software is quite costly and closed source and users should purchase separated modules as needed for the analysis. Command line-based approach allowed more freedom in applying specific types of analysis and thus better isolate comparisons, nevertheless bioinformatics skills are required.

3. Assessing optimal treatments for intracellular infection: host immunity, heterogeneity, and the antibiotic resistance challenge

Francisco F.S. Paupério[1,2], Erida Gjini[2]

1-Faculdade de Ciências da Universidade de Lisboa

2-Instituto Gulbenkian de Ciência

Mathematical models have been used as tools to study the dynamics of infectious diseases for a long time and to design successful control interventions, both at within-host and at the epidemiological level. Models can provide estimates of biological parameters which are difficult or expensive to obtain through experiments. Currently, in infection diseases, the growing antimicrobial resistance of pathogens poses great challenges. Recently, aggressive and moderate approaches are being debated as therapeutic strategies to deal with antibiotic resistance. In this work, we study intracellular infection dynamics combining effects of antibiotic treatment and adaptive immune responses. The ODE models are based on infection processes for acute and chronic bacterial infections. We find the critical parameter combination in macrophage-bacteria-immunity interaction, dividing regimes of clearance and persistence of infection. Moreover, we study the consequences of antimicrobial treatment on many infection measures, including duration, bacterial burden, pathology and resistance. We notice that different combination of treatment duration and antibiotic doses can lead to the same infection outcomes and that the same treatment can have different effects if applied early or later during the infection course. Moreover, treatment is not always beneficial, as longer durations often select more resistant bacteria. We compare short (3 days) versus long (7 days) treatment duration in-depth. Long treatment duration is overall more efficient, with higher infection resolution. However, there are regimes where short treatment is non-inferior or even superior. Our results highlight the potential of new targeted treatments of intracellular infection, with lower antibiotic doses and duration, combined with sufficient immune action. From this, we extract optimization principles for infection over a range of scenarios and we discuss future directions for the improvement of this area, namely the importance of infection and immunity biomarkers at treatment onset.

4. YEASTRACT+: a comparative genomics platform for the analysis of transcriptional regulatory networks in yeast species

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], Mafalda Cavalheiro [2,3], Melike Yilmaz [2,3], Romeu Viana [2,3], Isabel Sá-Correia [2,3], Miguel C. Teixeira [2,3] and 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

YEASTRACT (YEAst Search for Transcriptional Regulators And Consensus Tracking - http://yeastract.com) is a database for the analysis and prediction of transcription regulatory associations at the gene/genomic levels in *Saccharomyces cerevisiae*. For the last 12 years, regular updates ensured 163000 regulatory associations, based on more than 1600 bibliographic references [Teixeira, NAR 46:D348-D353 2018]. Provided bioinformatics tools enable the user to exploit the existing information to predict the TFs involved in the regulation of a gene or genome-wide transcriptional response, while ranking those TFs in order of their relative importance.

Additionally, the PathoYeastract database (http://pathoyeastract.org) was created to provide a resource for clinicians and biomedical scientists working with pathogenic yeasts [Monteiro, NAR 45:597-603 2017], describing more than 49000 regulatory associations of four species responsible for more than 90% of all detected candidiasis: *Candida glabrata* , *Candida albicans* , *Candida parapsilosis* and *Candida tropicalis*. Recently, NCYeastract (http://ncyeastract.org) was created for the study and analysis of non-conventional yeasts such as the food spoilage yeast *Zygosaccharomyces bailii*, highly tolerant to weak acids, with the short term prospect of considering three additional non-conventional yeast species of biotechnological interest. ProBioYeastract was also created to study the molecular basis of probiotic activity in pharmaceutical and food industries (two strains of *S.cerevisiae* var. boulardii : unique28 and biocodex).

The YEASTRACT+ platform currently under development, will combine the existing database information of all yeast species, 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. As a prospect, we intend to make use of this platform to facilitate the study of other relevant yeast species with available annotated genomes in GenBank. By processing their genomes, we will automatically generate a dedicated website for each species, providing predictive transcriptional regulation analysis and visualization tools, based on comparative genomics.

Funding: LISBOA-01-0145-FEDER-022231 - BioData.pt Research Infrastructure, EC H2020 grant 676559 ELIXIR-EXCELERATE and Plurianual UID/CEC/50021/2019.

Special Session: Companies

Ana Sofia Moreira

ALS Global

Frederico Carpinteiro

Adapttech

Patrícia Oliveira

Bioinf2Bio

Simão Soares

SilicoLife

If you've been wondering which bioinformatics companies exist in Portugal and what do they do, this is a great opportunity to enlighten yourself. Here, companies like ALS, adapttech, Bioinf2Bio and SILICOLIFE will briefly present themselves and their work.

Panel Discussion: Research in Bioinformatics

GenomePT

BioData.pt

Discussion about the current status of Bioinformatic research in Portugal from the point of view of some of the biggest Bioinformatic networks in the country, namely, BioData and GenomePT.