IX

BIOINFORMATICS OPEN DAYS
UNIVERSITY OF MINHO
2020

CONFERENCE BOOK
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>WELCOME MESSAGE</td>
<td>3</td>
</tr>
<tr>
<td>VENUE MAP / FACILITIES</td>
<td>4</td>
</tr>
<tr>
<td>COMMITTEE</td>
<td>5</td>
</tr>
<tr>
<td>SCIENTIFIC PROGRAMME</td>
<td>6</td>
</tr>
<tr>
<td>LECTURES</td>
<td>8</td>
</tr>
<tr>
<td>PITCH BOOTCAMP</td>
<td>11</td>
</tr>
<tr>
<td>ORAL ABSTRACTS</td>
<td>13</td>
</tr>
<tr>
<td>POSTER ABSTRACTS</td>
<td>24</td>
</tr>
</tbody>
</table>
Greetings,

On behalf of the Bioinformatic Open Days organization, it is my pleasure to invite all curious people, students, researchers and academics, interested to acquire more knowledge in Bioinformatics, to attend and participate in our event.

Bioinformatics has been significantly growing as an interdisciplinary technological and scientific field on its own right, providing enormous challenges and opportunities both in research and for companies. Following what is happening around the world, in Portugal, Bioinformatics have experienced an outstanding growth over the past few years. This is reflected academically, through the development of prestigious post-graduations, and in the economical/business sector with the establishment of new start-up companies with international connections.

Bioinformatics Open Days is a student-led initiative, first held at Universidade do Minho, Braga in 2012. It aims to promote the exchange of knowledge between students, teachers and researchers from the Bioinformatics and Computational Biology fields. This event will occur at Universidade do Minho (Campus Gualtar) on the 19th (workshops), 20th and 21st of February.

Every year, this event aims to describe at a glance the present and the future of Bioinformatics, nationally and internationally. Studies from several research institutions and projects from various companies will be presented.

In conclusion, we welcome all the participants to this event, and we hope that everyone has a great attendance.

The organizing committee of the BOD 2020
Ana Beatriz Gonçalves da Cunha
Ana Isabel Gomes Fernandes
Bruno Miguel Marques Pereira
Carolina Santiago Garrido Dias da Torre
Cátia Sofia Fonseca Gonçalves
Débora Alves Antunes
Maria Fernanda Silva Vieira
Gil da Lomba Afonso
Joana Raquel da Rocha Santos
João Manuel Capela Araújo Ribeiro
João Pedro Porto Dias
Maria Inês Alves Faria
Maria João Mogadouro Lopes
Nuno Miguel Caetano Alves
Pedro Henrique Matela Aidos Manso de Araújo
Pedro Daniel Martins Moreira
Raquel Sofia Vasconcelos Cardoso
Rui Ricardo Barros Nunes
Tiago Manuel Rocha Ferreira
Tiago Manuel Pereira de Oliveira
### WEDNESDAY 19th

<table>
<thead>
<tr>
<th>Time</th>
<th>Activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>10:00-13:00</td>
<td>“Studying Bioinformatics” Session – Lecture Room A1 (Department of Informatics)</td>
</tr>
<tr>
<td>13:00-14:00</td>
<td>Lunch</td>
</tr>
</tbody>
</table>
| 14:30-17:00| Workshop 1 – “Introduction to DNA barcoding bioinformatics” – 0.04 Department of Informatics  
Workshop 2 – “Single cell RNA-seq analysis using a seurat based graphical interface” – 0.12 Department of Informatics |
| 18:00      | Barbecue                                                                |

### THURSDAY 20th

<table>
<thead>
<tr>
<th>Time</th>
<th>Activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>09:30-10:00</td>
<td>Check-in</td>
</tr>
<tr>
<td>10:00-10:30</td>
<td>Opening Session</td>
</tr>
<tr>
<td>10:30-11:15</td>
<td>Keynote Lecture: Marwin Segler (BenevolentAI)</td>
</tr>
<tr>
<td>11:15-11:35</td>
<td>O1: Pedro Ribeiro - OphiDx: a bioinformatics platform for routine operation in a genetics laboratory</td>
</tr>
<tr>
<td>11:35-11:55</td>
<td>O2: Miguel Pinto – “WGS-based surveillance of Neisseria gonorrhoeae identifies major genogroups circulating in Europe associated with antimicrobial resistance profiles”</td>
</tr>
<tr>
<td>11:55-12:15</td>
<td>O3: Catarina Mendes - Dengue virus identification and genotyping from amplicon and shotgun metagenomic sequencing with DEN-IM</td>
</tr>
<tr>
<td>12:15-13:00</td>
<td>Lecture: Joel Arrais</td>
</tr>
<tr>
<td>13:00-14:00</td>
<td>Lunch</td>
</tr>
<tr>
<td>14:15-15:00</td>
<td>Lecture: João Saraiva</td>
</tr>
<tr>
<td>15:00-15:20</td>
<td>O4: Diogo Lima – “merlin v4: an updated platform for reconstructing genome-scale metabolic models”</td>
</tr>
<tr>
<td>15:20-15:40</td>
<td>O5: David Henriques – “Bioinformatics and systems biology for the design of novel wines”</td>
</tr>
<tr>
<td>15:40-16:00</td>
<td>O6: Jorge Comas – “A new genome-scale model predicts the growth of A. thaliana heterotrophic cell cultures in stress conditions”</td>
</tr>
<tr>
<td>16:00-16:45</td>
<td>Posters and Coffee Breaks</td>
</tr>
<tr>
<td>16:45-17:15</td>
<td>Lecture: Ahmad Zeidan</td>
</tr>
<tr>
<td>17:15-18:00</td>
<td>Debate: “Working abroad in bioinformatics”</td>
</tr>
<tr>
<td>18:30-20:00</td>
<td>Bioinformatics Quizz</td>
</tr>
<tr>
<td>20:30</td>
<td>Social Dinner</td>
</tr>
<tr>
<td>Time</td>
<td>Event Description</td>
</tr>
<tr>
<td>--------------</td>
<td>-----------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>10:00-10:30</td>
<td>Coffee Break</td>
</tr>
<tr>
<td>10:30-12:45</td>
<td>Companies’ Pitch Bootcamp</td>
</tr>
<tr>
<td>13:00-14:00</td>
<td>Lunch</td>
</tr>
<tr>
<td>14:00-15:10</td>
<td>Lecture: José Luís Oliveira</td>
</tr>
<tr>
<td>15:30-15:50</td>
<td>O8: Francisco Pina Martins – “Natural selection detection revisited – challenges and insights from automating and massively scaling the process”</td>
</tr>
<tr>
<td>15:50-16:10</td>
<td>O9: Rita Magalhães – “A computational approach for the identification of new inhibitors against biofilm set-up and development”</td>
</tr>
<tr>
<td>16:10-16:30</td>
<td>O10: João Cardoso – “Introducing the &quot;Ready for BioData Management?&quot; Programme”</td>
</tr>
<tr>
<td>17:00-17:25</td>
<td>Ending Session</td>
</tr>
</tbody>
</table>
Machine Learning for Molecular Design and Chemical Synthesis  
PhD Marwin Segler

Modern Machine Learning complements the way scientists can solve challenging problems, by allowing the machine to learn from data instead of manually writing a complex program for a specific task. In this talk, we will discuss how to apply ML in science in two examples:
First, we will discuss how neural networks can learn to generate structured data, and in particular molecules with desirable properties. This allows to automatically design molecules, for example for drug discovery.
Second, we will cover the grand challenge of automated synthesis planning, which can be used to predict how a given molecule can be synthesised in the laboratory. We will highlight recent progress on how to use deep neural networks to predict single step reactions, and how to combine machine learning with modern search techniques. We will close the discussion by addressing unsolved challenges, and the implications for drug design.

Computer-Guided Selection and Development of Functional Microbes  
PhD Ahmad Zeidan  
Computational Systems Biology, Discovery, R&D, Chr. Hansen A/S, 2970 Hørsholm, Denmark

Through continuous advances in sequencing technologies, microbial whole-genome sequences of increasing quality can be obtained at reasonable cost. While the availability of whole-genome sequences for many reference microorganisms has greatly benefited research and contributed to our overall understanding of microbial physiology, current technologies enable us to routinely sequence every individual strain. This opens up new avenues for life science research and development, including the discovery of the genetic basis underlying phenotype variability as well as in silico prediction of desirable phenotypes based on genomic content.

Here we will present two different approaches we apply to leverage the value of thousands of whole-genome sequences from the Chr. Hansen Culture Collection to guide the selection and development of Good Bacteria™, which can be used for the production of healthy food and/or contribute to improving human, plant and animal health. The first approach is a mechanistic approach, which relies on the use of strain-specific genome-scale models for unraveling the metabolic potential of our strains and guide the areas of strain development, culture development and bioprocess optimization. The second approach is a data-driven approach, where we develop machine learning models that can be used to accurately predict industrially relevant microbial traits and identify the underlying genetic determinants without a priori mechanistic knowledge. Application examples illustrating the potential of both approaches will be described.
The regular collection of Electronic Health Records (EHR) data in healthcare institutions have been creating an impressive health resource which potential value goes much beyond its primary use, either clinical or administrative. These databases can be used to speed up and reduce the cost of clinical research leading to advances and improvement of health services. At the same time, the increasing complexity of clinical studies, especially the ones with multifarious pathophysiology, demand for new processing approaches where omics' analysis are mixed with observational data to provide more precise treatments. However, many times, a single observational data source is not enough for a clinical study, thus data interoperability has a major impact on the exploration of value of EHRs. This talk will present and discuss some international initiatives that are dedicating effort to harmonize biomedical data so that it can be used in large research studies, maintaining ownership and privacy of local databases.

The advancements in high throughput sequencing technologies have greatly increased our ability to generate genomic data. This exponential increase in microbial big data now demands the development of novel concepts and strategies to better characterize and study microbial community composition and microbial interactions. The overall goal of João’s research is the development and implementation of novel concepts focused on microbial communities. Moreover, the main driver of his research activities relies on the generation of models that use (meta)genomic data to answer questions related to complex microbial communities and interspecies interactions. Because natural microbial communities are highly complex, answering these questions require careful design and planning to reflect its true scenario. In his research, he uses machine learning approaches that allow the discovery of patterns in microbial communities’ composition and interactions. In this talk will be presented the different projects that João Saraiva have worked on throughout his academic path, ranging from Metabolic network reconstruction, to machine learning in biomarker discovery to concept development in microbial ecology.
The traditional drug discovery process may take up to 15 years from conceptualization to market with a cost that can reach one thousand million, without any warranties that the identified compounds will reach the market. This is mainly a data-driven process that starts with all human proteins that can be used as putative targets, the millions of lead compounds that need to be evaluated and, for the final candidates, a massive number of structural variants to be tested. Deep networks were proven to be more effective than shallow architectures to face complex problems like speech or image recognition. In addition, deep architectures are able to amplify key discriminative aspects from the input data while suppressing irrelevant information, thus attaining improved accuracy. In this talk we present a walkthrough on the creation of a computational pipeline that uses Bioinformatics methods allied with Deep Learning architectures to support the drug discovery process.
Portugal’s Biotechnology Industry Organization (P-BIO) is the only association that brings together the vast majority of companies linked to the biotechnology and life sciences sector. Since it was founded in 1999, it has been the cornerstone for development and support of biotechnology in Portugal. P-BIO seeks to develop an environment that is favourable to the creation and growth of start-ups, promoting their corporate development domestically and internationally. While developing this ecosystem, it contributes for raising the profile of this sector and its developments. As a member of EuropaBio, the Organization is key to linking companies and their relevant partners in government, investors, regulating agencies and other institutions linked to the industry.

Chr. Hansen is a global bioscience company that develops natural solutions for the food, nutritional, pharmaceutical and agricultural industries. They develop and produce cultures, enzymes, probiotics and natural colors for a rich variety of foods, confectionery, beverages, dietary supplements, animal feed, and plant protection. Their product innovation is based on around 40,000 microbial strains – they like to refer to them as “good bacteria.” Their solutions enable food manufacturers to produce more with less – while also reducing the use of chemicals and other synthetic additives – which makes their products highly relevant in today’s world. They have been delivering value to their partners – and, ultimately, end consumers worldwide – for more than 145 years. We are proud that more than 1 billion people consume products containing our natural ingredients every day.

BioData.pt is the Portuguese distributed e-infrastructure for biological data and the Portuguese ELIXIR node. BioData.pt supports the national scientific system through best practices in data management and state of the art data analysis. It interfaces with both academia and industry, making research available for innovation, namely in sectors such as agro-food and forestry, sea, and health. BioData.pt services include ELIXIR services such as our training programme and computing facilities, as well as consulting services in data analysis and management, and a number of community services.

Founded in 2010, SilicoLife designs optimized microorganisms and novel pathways for industrial biotechnology applications. Based on metabolic engineering and synthetic biology approaches shortens the development time and costs of new highly effective processes for the production of specific target compounds such as chemicals, food ingredients or biopolymers.
Ophiomics offers innovative, added value, molecular diagnostic tests enabling a precision medicine, and supporting its services in an intensive Research & Development activity. Ophiomics aims to support the clinicians in implementing a precision medicine, allowing them to support their patients in choosing the best lifestyles tailored to their genetic background, in defining personalised strategies for early detection of disease, in implementing tailored therapies and in following therapeutical response. Their know-how and advanced infrastructure allows them to offer more and better information to the clinician, in time to be clinically useful. They focus on chronic diseases, including cancer. They offer tests based on DNA sequencing, where the starting sample is blood in the context of susceptibility, early detection and follow up studies, and tumor biopsies (fresh/FFPE) for pharmacogenomic tests.

PeekMed strive to build an unparalleled brand capable of developing engineering services for the community by creating innovative technological solutions which will improve the healthcare services, and increase the value of their clients, collaborators and partners. Their company aims to hold on the forefront of innovation, through the creation of products that redefine services in Healthcare.

Amyris

They believe highly-engineered organisms and sustainably sourced sugarcane can help solve world problems and support a healthier planet. That’s why they are using their technology to create products that support biopharmaceutical drug discovery and production, from cosmetic emollients and fragrances, to fuels, solvents, lubricants, and nutraceuticals. They believe it’s only a matter of time before everyone makes products their way. Make good. No compromise.
**Dengue virus identification and genotyping from amplicon and shotgun metagenomic sequencing with DEN-IM**

C I Mendes¹,², E Lizarazo², M P Machado¹, D N Silva¹, A Tami², M Ramirez¹, N Couto², J W A Rossen², J A Carriço¹

¹ Instituto de Microbiologia, Instituto de Medicina Molecular, Faculdade de Medicina, Universidade de Lisboa, Lisboa, Portugal; ² University of Groningen, University Medical Center Groningen, Department of Medical Microbiology and Infection Prevention, Groningen, The Netherlands;

Dengue virus (DENV) is the most clinically significant arbovirus worldwide. It is mainly concentrated in tropical and subtropical regions and represents a public health threat and economic burden in affected countries. The availability of genomic data is key to understanding viral evolution and dynamics, supporting improved control strategies. The use of High Throughput Sequencing (HTS) technologies, which can be applied both directly to patient samples (shotgun metagenomics) and PCR amplified viral sequences (amplicon sequencing), is potentially the most informative approach to monitor viral dissemination and genetic diversity by providing, in a single methodological step, identification and characterization of the whole viral genome at the nucleotide level.

We present DEN-IM, a one-stop, user-friendly, containerised and reproducible workflow for the analysis of DENV short-read sequencing data, both from amplicon and shotgun metagenomics approaches. DEN-IM is able to infer the DENV coding sequence (CDS), identify the serotype and genotype, and generate a phylogenetic tree, providing the results in an interactive HTML report that can be easily shared. Due to its Nextflow implementation and packaging of all dependencies as Docker containers, DEN-IM can be run on any UNIX-like system, from local machines to high-performance computing clusters, performing a comprehensive analysis without the requirement of extensive bioinformatics expertise. For DENV read identification, a database containing 3858 complete DENV genomes is provided alongside the workflow. For sero- and genotyping, we have curated and validated a selection of 161 complete DENV genomes representing the known diversity from the whole database through sequence identity and phylogenetic methods.

Using DEN-IM, we successfully analysed two types of DENV datasets. The first comprised 25 shotgun metagenomic sequencing samples of patients with variable serotype and genotype, including an in vitro spiked sample containing the four known serotypes. The second consisted of 106 paired-end and 76 single-end amplicon sequences of DENV 3 genotype III and DENV 1 genotype I, respectively, where DEN-IM allowed detection of the intra-genotype diversity. DEN-IM enabled the recovery, identification and typing of DENV with high accuracy from samples with as little as 1% DENV reads.

The DEN-IM workflow and documentation are freely available at [https://github.com/B-UMMI/DEN-IM](https://github.com/B-UMMI/DEN-IM).
Characterization of genomic variation in Portuguese sheep breeds using whole genome resequencing

D. Gaspar¹,², H. Magalhães¹, A. Usiê¹,³, C. Leão⁵, M. Monteiro⁴, M. Madeira⁴, J. Santos⁴, L. Tábuas⁴, S. Branco³, E. Bettencourt³, L. Padre³, R. Romão³, P. Caetano³, P. Damião³, C. Dias³, N. Carolino⁵, C. Bettencourt⁶, C. Ginja², C. Matos⁴, A.M. Ramos¹,³


Merino, Campaniça and Serra da Estrela (SE) sheep are among the most relevant breeds reared in Portugal. Merino and Campaniça sheep are mainly distributed in the south of Portugal, in the Alentejo region, being the basis for the production of different meat, dairy and wool products. SE is the main Portuguese dairy breed, located in the Serra da Estrela region. From its milk a typical, high-value cheese is produced, awarded with the protected designation of origin mark certification. Despite their importance, the lack of significant genomic resources is a problem shared by these breeds. Thus, the purpose of the present study was to assess the variability present in the genomes of these sheep breeds, and a population of crossed Merino sheep, using whole-genome resequencing (WGRS).

A total of 34,976,564,162 paired-end raw reads were produced for 56 sheep samples, from which, 93.2% were kept for downstream analysis after quality control procedures. With a mapping rate of around 99.85%, an average of 90.01% of high-quality reads were uniquely mapped to the sheep reference genome. Variant calling was performed and an initial set of 115,137,724 raw SNPs was obtained. After SNP filtering, a final set of 31,320,381 high-quality SNPs were maintained. A total of 11,148,321 SNPs were located in genic regions, where 120,172 were annotated as synonymous and 80,882 as non-synonymous. Moreover, 20,172,060 SNPs were identified in intergenic regions. Lastly, structural variation was also characterized in all 56 sheep genomes.

The results derived from this study will be useful to develop several genomic tools for these breeds, including genome-wide association studies, genetic diversity and traceability schemes.
Bioinformatics and systems biology for the design of novel wines

D. Henriques¹, R. Minebois², L. G. Macías³, R. Pérez-Torrado², E. Barrio³, A. Querol, E. Balsa-Canto¹.


Abstract:

Wine is an alcoholic beverage with great cultural and economic importance which results from the conversion of grape must. Modern wine-production uses selected yeasts, as starters to control the fermentation, reducing bacterial contamination, increasing reproducibility and generating wines with specific characteristics. However, the new challenges of the winemaking industry, namely the reduction of wine alcohol levels, the production of distinctive wines with improved aroma profiles and the reduction of energy use, call for the use of new starters. The selection of one specific species will depend on the grape cultivar, must composition, the final product requirements and, most importantly, their ability to adapt to winemaking conditions. Non-cerevisiae species of the Saccharomyces genus have shown potential but is critical to gain further knowledge about their metabolism. For that purpose, we suggest using a model-based systems biology approach.

Modelling yeast metabolism in winemaking conditions is challenging because of the grape must complexity and the dynamic nature of the process. During fermentation, yeasts suffer rapid nutrient limitation and ethanol toxicity. To account for this complexity, we combined 1) bioinformatics tools to adapt Yeast8 reconstruction to wine conditions, 2) time-series data of relevant metabolites from micro-vinification experiments driven by various yeast species and 3) an identification procedure within the dFBA (dynamic flux balance analysis) framework.

The final model predicts the dynamics of the production of aromas while being compliant with all other measured external metabolites and biomass formation. We found that S. uvarum appears to operate both a reductive (cytosolic) and oxidative (mitochondrial) branch of the TCA cycle which explains the shift of carbon from ethanol to other products along with its acetate consumption and succinate production during stationary phase. Additionally, we found that considering protein turnover and using a dynamic biomass equation are critical for modelling wine fermentations.
**merlin v4: an updated platform for reconstructing genome-scale metabolic models**

Diogo Lima¹, Davide Lagoa¹, Fernando Cruz¹, José Bastos¹, João Capela¹, Eugénio Ferreira¹, Miguel Rocha¹ and Oscar Dias¹

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

The Metabolic Models Reconstruction Using Genome-Scale Information (merlin) software (1) is an open source user-friendly Java (2) application developed for Windows and Unix, aimed towards the reconstruction of genome-scale metabolic models. The development of merlin follows a design philosophy of automating time-consuming steps in the reconstruction of genome-scale metabolic models, while allowing users to control the parameters of operations and manually curate the results. All major steps involved in the reconstruction of a metabolic model are implemented in merlin, including genome retrieval and its functional annotation, construction of the reactions’ set and associated entities, model compartmentalization and conversion to standard SBML formats (3).

The fourth iteration of merlin includes a major overhaul of the user interface, implementation of new features, improvements to existing features, and most notably, the implementation of the object-relational mapping framework Hibernate (4). The graphical layout has been significantly streamlined, while supporting the latest version of AiBench (5), providing users with an intuitive and responsive interface. Development was also focused at new quality of life improvements, aimed mainly towards importing, exporting and duplicating merlin user projects. The development of the latest version of merlin followed a modular approach, culminating in the implementation of a plugin manager which simplifies and hastens the process of updating and debugging the various features of merlin.

In addition, TranSyT, a state-of-the-art genome-wide transmembrane transport system annotation tool has been implemented to overcome the limitations of the previously available TRIAGE module (6). Finally, it is noteworthy to mention the implementation of BioISO, a tool aimed at evaluating a genome-scale metabolic network or biomass formulation, based on the previously available COBRA (7) and FBA (8) frameworks.
Natural selection detection revisited – challenges and insights from automating and massively scaling the process

F. Pina-Martins\textsuperscript{a}, S. Andaluz\textsuperscript{bc}, M. Janeiro\textsuperscript{bd}, B. Monteiro\textsuperscript{be}, O.S. Paulo\textsuperscript{b}

\textsuperscript{a}Computational Biology and Population Genomics Group, Centre for Ecology, Evolution and Environmental Changes, Departamento de Biologia Animal, Faculdade de Ciências, Universidade de Lisboa, Campo Grande, 1749-016 Lisboa, Portugal
\textsuperscript{b}Departamento de Biologia Animal (DBA), Faculdade de Ciências, Universidade de Lisboa, Campo Grande, 1749-016 Lisboa, Portugal
\textsuperscript{c}Laboratório de Saúde e Produção Animal Tropical (TAHP), CIISA - Centro de Investigação Interdisciplinar em Sanidade Animal, Faculdade de Medicina Veterinária, Universidade de Lisboa, Avenida da Universidade Técnica, 1300-477 Lisboa, Portugal
\textsuperscript{d}Laboratório de Genética e Instituto de Saúde Ambiental, Faculdade de Medicina da Universidade de Lisboa, Av. Prof. Egas Moniz, Piso 1C | 1649-028 Lisboa
\textsuperscript{e}CESAM - Centro de Estudos do Ambiente e do Mar, Universidade de Aveiro Campus Universitário de Santiago 3810-193 Aveiro, Portugal

The detection of genetic variants putatively under selection is an approach that can provide insights into local adaptation events. The number of studies employing such approaches has been steadily increasing through time, especially as High Throughput Sequencing methods gain adoption among the scientific community.

Current methods for performing these detections are of essentially two types – “Outlier Analysis” (OA) and “Environmental Association Analysis” (EAA). Within each of these categories, there are currently multiple methodologies and implementations, but there is still no community standard way to perform such analyses, and therefore, no reliable way to compare results.

Past efforts, have attempted to summarize the effectiveness of such approaches by gathering large number of studies from the literature and comparing their results. Conclusions reached by this approach, however, are limited due to the diversity of methodologies and scopes used to perform selection detection.

In this work we are mining the literature for papers that performed either OA or EEA that have a) made their data available b) use SNP markers, and c) samples are GEO-referenced. Unlike in previous works, the data gathered here is then re-analysed always using the same methods, Bayescan and SelEstim for OA and Baypass and LFMM for EEA.

In order to scale up to hundreds of datasets and remain reproducible, the entire pipeline is automated via GNU Make and run on container technology (currently docker).

Obtained results are then compared by category of organism and geographic scope and with those of previous studies.
Introducing the "Ready for BioData Management?" Programme

João Cardoso¹, Daniel Faria¹, José Borbinha¹

¹BioData.pt & INESC-ID, Lisboa, Portugal

Abstract

Data management is a key issue in modern science, as the rate of data production demands efficient solutions for finding and interpreting data and requires that data be published in a manner amenable to such solutions. This requirement has been formalized as the FAIR data principles of Findability, Accessibility, Interoperability and Reusability, which are being promoted by ELIXIR, the European infrastructure for data in the life sciences. Funding agencies increasingly require the projects their fund to adopt Open Science policies and abide by the FAIR data principles. Furthermore, most funding agencies now require projects to include data management plans (DMPs). However, few researchers are aware of these needs and requirements or prepared to create a suitable DMP for a research project.

BioData.pt¹, the Portuguese node of ELIXIR, aims to bridge this gap and support the Portuguese research community by providing training and consulting services in data management. These services were kicked-off in 2019 with the launching of the "Ready for BioData Management?"² Programme. The first event of this programme aimed at raising awareness and demystifying DMPs, through an innovative interactive hands-on group exercise where participants filled-in a DMP for a mock project on a large canvas. This basic recipe was well-received by participants, and showed that the topic is in-demand, leading BioData.pt to amplify its offer.

Currently the "Ready for BioData Management?" programme includes the following event formats: (1) A one-day introductory workshop about DMPs following the recipe of the first event; (2) A one-day advanced bring-your-own-data DMP course where will be guided through the creation of their own DMP, using a software framework; (3) Class modules aimed at graduate students introducing data management and DMPs.

A second introductory workshop and class modules were already held in 2019, and five "Ready for BioData Management?" events spanning all three formats are scheduled for the first trimester of 2020, which clearly shows that the programme is satisfying a critical need of the Portuguese research community. For 2020, we plan on expanding the offers of the programme further by tackling the issue of day-to-day research data management, with both an introductory one-day workshop and an in-depth week-long course.

Acknowledgments

This work was supported by national funds through Fundação para a Ciência e a Tecnologia (FCT) with reference UID/CEC/50021/2019, and project ELIXIR EXCELERATE.
A new genome-scale model predicts the growth of *A. thaliana* heterotrophic cell cultures in stress conditions

Jorge Comas<sup>1*</sup>, Nelson Saibo<sup>1</sup>, Claudine Chaouiya<sup>2,3</sup>, Margarida Oliveira<sup>1</sup>

1- ITQB NOVA, Av. da República, 2780-157 Oeiras, Portugal
2- IGC, Rua da Quinta Grande 6, 2780-157 Oeiras, Portugal
3- Aix Marseille Univ, CNRS, Centrale Marseille, I2M, Marseille, France
* jorgecp@itqb.unl.pt

Genome-scale metabolic models and constraint-based analysis tools have been used in the past two decades to predict metabolic fluxes, first in prokaryotes [1], and subsequently in eukaryotes [2], at the quantitative level. Recently, several genome-scale models of the model plant *A. thaliana* have been constructed using various approaches and made available in the literature [3, 4, 5]. Our goal is to use similar approaches to elucidate differences in the metabolic flux phenotypes characteristic of plant growth in different stress conditions, namely in osmotic stress and salt stress.

To approach this issue, we developed a new genome-scale model of *A. thaliana* that combines the best features of the available models, including high coverage of reactions and high confidence in gene-protein-reaction associations. This new model has better predictive capabilities regarding fluxes in the central carbon metabolism evaluated against metabolic flux analysis data obtained in various stress conditions. Additionally, we demonstrate the ability of the new model to accurately predict the growth of *A. thaliana* heterotrophic cell cultures over time, in optimal conditions, and when subjected to osmotic stress and salt stress.

Our results indicate that the constraint-based approach is suitable to elucidate the dynamics of metabolic flux phenotypes of plant cells and to study the main features of growth under stress conditions at the metabolic flux level.
WGS-based surveillance of Neisseria gonorrhoeae identifies major genogroups circulating in Europe associated with antimicrobial resistance profiles

Miguel Pinto¹, Vitor Borges¹, Joana Isidro¹, Luís Vieira³⁴, Maria José Borrego², João Paulo Gomes¹

¹Bioinformatics Unit, Department of Infectious Diseases, National Institute of Health, 1649-016 Lisbon, Portugal
²Reference Laboratory of Bacterial Sexually Transmitted Infections, Department of Infectious Diseases, National Institute of Health, 1649-016 Lisbon, Portugal,
³Technology and Innovation Unit, Department of Human Genetics, National Institute of Health, Lisbon, Portugal.
⁴Centre for Toxicogenomics and Human Health (ToxOmics), Genetics, Oncology and Human Toxicology, Nova Medical School/Faculty of Medical Sciences, New University of Lisbon

Neisseria gonorrhoeae, the bacterium responsible for the sexually transmitted disease gonorrhea, has shown an extraordinary ability to develop antimicrobial resistance (AMR) to multiple classes of antimicrobials. The observed steady rise in AMR and disease burden remain major public health concerns worldwide [1, 2]. With no available vaccine, managing N. gonorrhoeae infections still relies on preventive measures, effective antibiotic treatment and epidemiological surveillance. This has been the focus of worldwide national surveillance programmes such as GRASP, EURO-GASP and GISP, which are now transitioning to a whole-genome sequencing (WGS)-based surveillance.

The present study, focused in Europe, aims to classify N. gonorrhoeae into genogroups by performing isolate clustering based on WGS data for prospective WGS-based surveillance. We aimed to identify the major circulating WGS-genogroups, to establish a relationship between these and AMR, and to contribute with WGS data from Portuguese isolates spanning 15 years of surveillance.

A total of 3791 carefully inspected N. gonorrhoeae genomes from isolates collected across Europe (assembled with the INNUca pipeline [3]) were analyzed using a gene-by-gene approach (chewBBACA [4]). After clustering using goeBURST [5], cluster composition was compared at all possible allelic distance thresholds. Two distinct cluster stability points were identified, allowing the classification of isolates into a higher and lower WGS-based Genogroup.

Data revealed the existence of 180 Higher-Genogroups and 321 Lower-Genogroups. Thirty-eight of the latter are composed by ≥ 10 isolates each, which comprise 84.4% of the whole dataset analyzed. Results show a relationship between the major genogroups circulating and carriage of genetic determinants of AMR and decreased susceptibility, namely for penicillin, tetracycline, ciprofloxacin, cephalosporins and azithromycin (the two latter being the current recommend empirical treatment options). Analysis at a lower threshold level within each Lower-Genogroup also allows us to relate clustering data with previously described outbreaks.

Globally, this study: i) reveals the major N. gonorrhoeae WGS-based genogroups that should be monitored and ii) demonstrates the usefulness of continuous characterization of circulating strains and AMR emergence monitoring for prospective WGS-based surveillance and outbreak detection.
OphiDx: a bioinformatics platform for routine operation in a genetics laboratory

Authors:

Pedro Ribeiro², Ana Pereira², Joana Cardoso¹, José Pereira-Leal¹, José Germano de Sousa², Maria José Rego de Sousa², José Germano Rego de Sousa²

In recent years genomic approaches have gained increasing acceptance in the clinical setting as the costs of Next Generation Sequencing (NGS) equipment and of sequencing reagents both steadily decreased. It is becoming increasingly common to sequence large gene panels and exomes in the germ line setting, as well as large somatic gene panels in solid tumours. This creates new challenges for clinical laboratories that have to manage and make sense of the increasing amounts of data from NGS, as well as data coming from classical sources that are gradually automated, for example genotyping. The goal of the clinical genetics laboratory is to analyze the genetic information of a sample, and to provide a characterization that is clinically meaningful and actionable, which is becoming increasingly complex due to the data deluge mentioned above. New artificial intelligence platforms, most notably IBM’s Watson for Genomics (WfG), can assist laboratory and clinical geneticist to make sense of this data, however they need to be integrated into a data pipeline. At the Centro de Medicina Laboratorial Germano de Sousa, we have been developing an in-house bioinformatics platform to integrate multiple data types, external sources as WfG, automating as many steps of data analyses as is feasible and simplifying the life of laboratory geneticists that deal with hundreds of life or death situation every month. OphiDx was built on a background of Django and PostgreSQL and powers the operation of multiple test types, including NGS, integrate recommendations from WfG and output a clinical report in PDF and/or HL7 format. I will discuss the challenges of deploying a bioinformatics platform for routine operation and illustrate its use with a description of the deployment of a completely automated data analyses and reporting pipeline for gut microbiomes recently developed at our

1. Ophiomics - Precision Medicine, Lisboa, Portugal
2. Centro de Medicina Laboratorial Germano de Sousa, Lisboa, Portugal
3. iNOVA4Health-Advancing Precision Medicine, NOVA University of Lisbon, Lisboa, Portugal
4. NOVA Medical School, FCM, Lisboa, Portugal
Chewie-NS: Enabling the use of gene-by-gene typing methods through a public and centralized service

Pedro Vila-Cerqueira¹, Rafael Mamede¹, Mickael Silva¹, João André Carriço¹, Mário Ramirez¹

¹Instituto de Microbiologia, Instituto de Medicina Molecular, Faculdade de Medicina, Universidade de Lisboa, Lisboa, Portugal.

The recent advances in High Throughput Sequencing (HTS) technologies led to substantial decreases in costs associated with microbial genomics approaches, favoring the transition from traditional Multilocus Sequence Typing (MLST) to core-genome and whole-genome based approaches (cg/wgMLST).

We have previously proposed a suite, chewBBACA, that allows the creation of gene-by-gene schemas and determination of allelic profiles from assembled draft genomes. Nevertheless, allelic profiles should be shareable and comparable within the community at a global level. For that purpose we developed Chewie-NS, a Nomenclature Server that is based on the TypOn ontology and integrates with the chewBBACA suite to provide a centralized service to download or update schemas, allowing the easy sharing of results, while ensuring the reproducibility and consistency of these steps.

Chewie-NS provides a public and centralised web service, with a backend built with Flask, a Python web development microframework, and uses the Virtuoso triple store as the main database. The frontend was built using the ReactJS web framework, providing a user friendly interface. Chewie-NS is an easy way for users worldwide to download the necessary data defining the cg/wgMLST schemas, perform analyses locally with chewBBACA, and, if they so wish, to submit their novel results to the web service through a REST API to ensure that a common nomenclature is maintained.

Our approach is distinct from other publicly available web services, such as PubMLST or Enterobase, by letting the users run their analyses in local machines, circumventing possible concerns on data privacy. This is reinforced by the possibility to maintain novel alleles private. The deployment of local instances of Chewie-NS can be easily achieved through the use of Docker Compose. Chewie-NS allows users to share their data and compare it with the community while ensuring the reproducibility of their analyses and the privacy of the data.
A computational approach for the identification of new inhibitors against biofilm set-up and development

Rita P. Magalhães¹,*, Tatiana F. Vieira¹, André Melo² and Sérgio F. Sousa¹

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

Biofilms are highly organized communities of bacteria attached and enclosed in a self-produced matrix. These structures show differences in gene expression when compared with similar free-flowing cells. Furthermore, they are highly resistant to antibiotics and host immune response and affect both human tissues and medical devices.¹ *P. aeruginosa* is a highly pathogenic biofilm forming gram negative bacteria. The development of potent inhibitors against its mechanisms of biofilm formation is a promising therapeutic strategy to combat *P. a* related infection.²

Molecular Docking³ is a computational method used to accurately predict the preferred binding pose between two molecules. Virtual Screening⁴ (VS) is the application of docking to large databases of compounds. The development and optimization of specific VS protocols capable of identifying compounds with inhibition potential against LasR is a possible strategy to reduce biofilm formation in *P. aeruginosa*.

A comprehensive structural database of all known structures of proteins involved in biofilm formation and development, including biological, kinetic, mutagenic and inhibition data has been developed throughout this project. This work also reports on the identification of compounds for the inhibition of quorum-sensing in *P. aeruginosa*. Several molecular docking software were used in the screening of large of compounds with unknown activity against LasR. The best performing compounds can now be tested experimentally for their action against LasR.

Acknowledgements This work has been supported by the Fundação para a Ciência e a Tecnologia (FCT) (UID/Multi/04378/2019)
Universal laws for protein complementation as a therapeutic

Ana Catarina Narciso, Ricardo Letra-Vilela, Fernanda Murtinheira, Federico Herrera, Manuel Nuno Melo

Instituto de Tecnologia Quimica e Biologica (ITQB-NOVA), Universidade Nova de Lisboa, Oeiras, Portugal

Protein complementation (PC) is a property of most proteins that was first reported in 1958, in a pancreatic ribonuclease. It consists on the proteins’ ability to be split into two or more non-functional fragments, which are able to recover function once they are brought together. After this pioneer work, the same property was identified in a high number of proteins with technical interest, such as β-galactosidase, ubiquitin, thymidine kinase or the green fluorescent protein family.

However PC cannot, currently, be predicted a priori and needs to be identified experimentally - both time-consuming and expensive. More than sixty years after being discovered, PC is widely used as a technique to visualize or study protein-protein interactions. Nevertheless, the basic principles behind PC remain unknown and may have other applications beyond the analysis of protein-protein interactions.

Assuming all truncated proteins could be reassembled and giving that many human diseases are caused by deletions/truncations/mis cleavage of proteins, we suggest fragments can be designed to complement structure and rescue the function of disease-related proteins.

Using the well described PC in proteins from the green fluorescent protein (GFP) family as a starting point, we hope to unravel a universal model that allows the prediction of PC combinations in human health benefit. We focus our work on trying to create a computational model that integrates many protein fragments (the fluorescence of which depends) on a successful complementation. This model will predict which fragments can have a productive complementation a priori. To this end, we will feed our model with as many chemical-physical/structural/sequence features as possible: among others, fragment sequence overlap, distance of cleavage site to the fluorophore, structural exposure of the fluorophore after cleavage, solvent accessible surface area of the fragments, and hydrophobic/hydrophilic character of the cleaved interfaces.

Experimental work will be used to test our model in vivo. We expect to lay the groundwork for the discovery of universal rules of PC and thus enable the rational design of new complementation therapies for a wide spectrum of human pathologies.
Development of novel RdRP inhibitors to the Rabies virus: an in silico approach

Ana Filipa Mendonça, Ana Oliveira, Maria João Ramos, Pedro Fernandes
Departamento de Química e Bioquímica, Faculdade de Ciências da Universidade do Porto

Rabies disease is a global problem. Despite the decline of cases, due to vaccination campaigns, it is estimated that every year thousands of people die infected by Rabies virus. To date, the only treatment is the vaccination, and no antiviral drug is available on the market.

The Rabies virus is a negative-sense single stranded RNA which, similar to its virus class, express a RNA dependent RNA protein (RdRP) polymerase which catalyses the RNA replication.
To prevent the virus replication, and thus impede the disease, we aim to inhibit the RdRP enzyme by either locking the enzyme in an inactive conformation or prevent its binding to the RNA chain.

Very limited information on the 3D structure of the RdRP of Rabies Virus is known. The complex has an L-protein that copies RNA and a cofactor P that stimulates RdRP activity. RdRP is assembled in 3 domains: C-Terminal that binds the RNA, the oligomeration domain, and N-terminal that binds to the L-protein.

The first step of this work was to create a model of the RdRP using homology modelling techniques. The principle behind this technique is that proteins that display sequence similarity, tend to fold in similar 3D structures. The templates used to estimate our model were the L-protein of vesicular stomatitis virus, the influenza B polymerase, the reovirus type 3 and rotavirus polymerase VP1, PDB IDs. 5A22, 4WRT, 1MUK and 2R7Q, respectively. The obtained model was then evaluated by a DOPE score.

The next step is to identify druggable cavities and compounds that inhibit the RdRP enzyme. Therefore, we will perform a virtual screening of 3.5M commercially available compounds, that obey the lipinsky rules, using the rDock, Autodock vina and Gold software. The top ranked compounds will then be evaluated by an in vitro screening assay, in order to validate their biological activity, at Prof. Choowongkomon lab at the University of Kasertsart, in Thailand.
ProPythia, an automated platform for the classification of peptides/proteins using machine learning

Ana Marta Sequeira, Sara Pereira, Diana Lousa, Miguel Rocha

Authors affiliation: University of Minho, ITQB

One of the most challenging problems in bioinformatics is to computationally characterize sequences, structures and functions of proteins. Sequence-derived structural and physicochemical properties of proteins have been used in the development of machine learning models in protein related problems. However, tools and platforms to calculate features and perform Machine learning (ML) with proteins are scarce and have their limitations in terms of effectiveness, user-friendliness and applicability.

Here, a generic modular automated ML-based platform for the classification of proteins based on their physicochemical properties is proposed. ProPythia, developed as a Python package, facilitates the major tasks of ML and includes modules to read and alter sequences, calculate protein features, pre-process datasets, execute feature reduction and selection, perform clustering, train and optimize ML models and make predictions. This platform was validated by testing its ability to classify anticancer and antimicrobial peptides and further used to explore viral fusion peptides.

Membrane-interacting peptides play a crucial role in several biological processes. Fusion peptides are a subclass found in enveloped viruses, that are particularly relevant for membrane fusion. Determining what are the properties that characterize fusion peptides and distinguishing them from other proteins is a very relevant scientific question with important technological implications.

Using three different datasets composed by well annotated sequences, different feature extraction techniques and feature selection methods, ML models were trained, tested and used to predict the location of a known fusion peptide in a protein sequence from the Dengue virus. Feature importance was also analysed. The models obtained will be useful in future research, also providing a biological insight into the distinctive physicochemical characteristics of fusion peptides.

This work presents a freely available tool to perform ML-based protein classification and the first global analysis and prediction of viral fusion peptides using ML, reinforcing the usability and importance of ML in protein classification problems.

Keywords: Machine Learning; Peptide Classification; Viral Fusion Peptides
Whole genome resequencing analysis of Alentejano pigs reveals differences associated with meat quality phenotypes

A. Usié1,2, H. Magalhães1, M. Antunes1, C. Leão1, D. Gaspar1, B. Meireles1, P. Barbosa1, L. Cacheu1, A. Albuquerque2, R. Charneca4, J. Martins3, E. Jerónimo1,2, A. M. Ramos1,2

1 Centro de Biotecnologia Agrícola e Agro-alimentar do Alentejo (CEBAL) / Instituto Politécnico de Beja (IPBeja), 7801-908 Beja, Portugal
2 MED – Mediterranean Institute for Agriculture, Environment and Development, CEBAL — Centro de Biotecnologia Agrícola e Agro-Alimentar do Alentejo, 7801-908 Beja, Portugal
3 MED – Mediterranean Institute for Agriculture, Environment and Development, Departamento de Zootecnia, Universidade de Évora, Évora, Portugal
4 MED – Mediterranean Institute for Agriculture, Environment and Development, Departamento de Medicina Veterinária, Universidade de Évora, Évora, Portugal

The Alentejano is a Mediterranean pig breed, found in southern Portugal, reared under extensive conditions and finished on grass and acorns during the fall and winter months. From these animals a variety of dry-cured meat products of great economic importance are generated. A total of 541 pigs were studied during the 2017 slaughter campaign. Phenotypic records for carcass and meat quality were collected and subsequently analyzed to identify the groups of animals that displayed the most contrasting phenotypes. Two groups comprising 13 animals each were selected, based on pH, water loss, total lipids, total protein, total collagen and pigments content. All samples were re-sequenced to a 23x coverage. The reads were aligned to the pig genome and SNPs and structural variants identified between the two groups of animals. A total of 13,418,254 SNPs were identified, of which 6,851,475 and 6,566,779 were located in the genic and intergenic genomic regions, respectively. The number of SNPs for which at least 25 samples were present comprised 88.7%. The set of genic SNPs included 43,405 exonic non-synonymous SNPs and 60,750 exonic synonymous SNPs. The remaining SNPs were located in introns and ncRNA regions. Interestingly, SNPs with markedly different allele frequencies between the groups were also identified (a total of 230 SNPs with allele frequency differences between the groups of at least 30% or 70%). This study represents the first major characterization of Alentejano pigs at the genome level, and identified a significant number of SNPs potentially associated with meat quality.
Dynamic modeling of the shikimate pathway and central carbon metabolism

André Fonseca[a], Isabel Rocha [b]

Instituto de Tecnologia Química e Biológica António Xavier (ITQB) of Universidade Nova de Lisboa (UNL) – afonseca@itqb.unl.pt [a], irocha@unl.pt [b]

The shikimate pathway is the central carbon metabolism’s (CCM) gateway to aromatic amino acids and other high value compounds important in industry and pharmaceuticals[1][2]. The biological synthesis of these metabolites is highly sought after due to its sustainability[2], however, the commercial viability of large-scale fermentation processes depends enormously on the microorganism's efficiency of producing the desired products[3]. While the constant development of metabolic engineering (ME) has been continuously increasing the widespread use of such processes, the sheer complexity of metabolic regulation makes it difficult to predict the ideal targets for ME[4]. To tackle this issue, mathematical models that can describe the biological systems behavior have become an essential ally[5]. These models are able to quantitively predict intracellular concentrations of metabolites, as well as fluxes, from ordinary differential equations based on kinetic equations that depict enzymatic reactions[5][6]. Because of the many biochemical details of metabolic networks, along with a lack of kinetic information on the dynamics of reactions, there are no dynamic models that encompass a cell’s entire metabolism, being more objective focused[7][8][9]. In this study, we present a dynamic model for Escherichia coli (E. coli) that simulates the CCM: glycolysis, pentose phosphate pathway, citric acid cycle (TCA), anaplerotic reactions and the glyoxylate pathway, along with the shikimate pathway and aromatic amino acids synthesis pathway. Fluxomic and metabolomic data was used to validate the model’s behavior, with simulations of wild-type and genetically modified strains (single knockouts). The results we obtained indicate that the present model can simulate the metabolic fluxes of E. coli and could therefore be a great tool for studying the control of such fluxes.
Characterization of the stone pine (*Pinus pinea*) needle transcriptome: *de novo* assembly and SNP identification

B.Mendes\(^1\)*, M.Antunes\(^1\)*, A.Usié\(^{1,2}\), C.Leão\(^3\), A.M.Ramos\(^{1,2}\)

\(^1\)Centro de Biotecnologia Agrícola e Agro-Alimentar do Alentejo (CEBAL) Instituto Politécnico de Beja (IPBeja), 7801-908 Beja, Portugal.\(^2\) MED-Mediterranean Institute for Agriculture, Environment and Development, Évora, Portugal.\(^3\) INIAV - Instituto Nacional de Investigação Agrária e Veterinária, Oeiras 2780-157, Portugal.

*These authors contributed equally to this work

*Pinus pinea*, commonly known as stone pine, is a native species widespread throughout the Mediterranean region. It is mainly associated with pine nuts production, commercialized in food industry, and also with timber. Stone pine trees are used as an ornamental species because of their unique features, being a very emblematic presence in many European countries and cities.

Despite its relevance, the volume of genomic information available for stone pine is very scarce. Since the number of studies and data available for this species are very limited, high-throughput sequence technologies were employed to generate a large scale RNA-seq dataset. The goal of this study was to characterize the stone pine needle transcriptome, by performing a *de novo* assembly, followed by SNP identification.

Needle samples were collected in 26 trees, from a population located in Coruche, Portugal. After RNA extraction and sequencing, the raw RNA-seq data were filtered by quality and length and used for *de novo* assembly and SNP calling. The transcriptome assembly was performed with the reads from five individuals previously selected using the Mira assembler, and then functionally annotated. Alignment of reads from 26 individuals was done against the needle transcriptome assembly previously generated using STAR. SNP calling was performed with freebayes and quality filtering was done using vcftools. Annovar was used to do the functional annotation of the SNPs. This work allowed the first characterization of the stone pine needle transcriptome, along with the SNP identification in the species.
Kinship analysis and Pedigree Reconstruction of a Self-Regenerating Cork Oak Population

B. Mendes¹, T. Sampaio², M. Antunes¹, H. Magalhães¹, A. Usié¹,⁴, F. Simões³, H. Almeida², A. M. Ramos¹,⁴

¹Centro de Biotecnologia Agrícola e Agro-Alimentar do Alentejo (CEBAL) Instituto Politécnico de Beja (IPBeja), Beja, Portugal. ²Instituto Superior de Agronomia (ISA) Instituto Superior de Agronomia, Lisboa, Portugal ³Instituto Nacional de Investigação Agrária e Veterinária (INIAV). ⁴MED-Mediterranean Institute for Agriculture, Environment and Development, Évora, Portugal

Cork oak (Quercus suber) trees are very important in the Mediterranean region, and are therefore the target of many studies. The measurements of the relatedness between individuals can be used to predict how a specific trait can pass through generations, producing information that can be used to improve the conservation and breeding programs of the species. So far, no studies about kinship and pedigree studies are available for cork oak. Thus, in this study a natural cork oak tree population was genotyped using a set of single nucleotide polymorphisms (SNPs), to predict and reconstruct its pedigree.

A total of 500 trees, from a natural cork oak population located in Herdade da Gâmbia (Portugal), were genotyped with 8,412 SNPs following a standard genotyping protocol. After SNP calling, the raw set of SNPs was filtered in four different ways, producing 4 SNP sets. SNPs with low quality and low coverage were removed in all sets. Each set was further filtered by different values of missing data, genotype frequency and minor allele frequency. Then, the identity by descent (IBD) matrix was generated in order to perform the relationship prediction. Among all sets, the number of relationships ranged from 22,345 to 24,729, with a total of 10,969 commonly found between all sets. Familial categories from the 1ˢᵗ to the 3ʳᵈ degree and monozygotic twins were assigned to most of the identified relationships. Because of the significant data complexity, full reconstruction of the pedigree for the whole population was not finalized. Different subsets were then created by maintaining only the trees related in first and second degree. The successful identification of kinships and establishment of pedigrees for this small families indicates the potential of this approach for future similar studies even though full pedigree reconstruction may prove to be difficult in large populations with convoluted familial relationships.
Sparse models and network-based regularizers for the analysis of RNA-seq data from colorectal cancer

Carolina Peixoto (anacpeixoto@tecnico.ulisboa.pt)¹, André Veríssimo¹ (andre.verissimo@tecnico.ulisboa.pt), Marta Martins² (marta.martins@medicina.ulisboa.pt), Marta B. Lopes³ (marta.lopes@tecnico.ulisboa.pt), Luís Costa (lmcosta@medicina.ulisboa.pt)²,⁴, Susana Vinga (susanavinga@tecnico.ulisboa.pt)¹

1. INESC-ID, Instituto Superior Técnico, Universidade de Lisboa, Rua Alves Redol 9, 1000-029 Lisboa, Portugal
2. Instituto de Medicina Molecular- João Lobo Antunes, Faculdade de Medicina de Lisboa, Avenida Professor Egas Moniz, 1649-028 Lisboa, Portugal
3. Instituto de Telecomunicações, Instituto Superior Técnico, Universidade de Lisboa, Av. Rovisco Pais 1, Lisboa, 1049-001, Portugal
4. Oncology Division, Hospital de Santa Maria, Centro Hospitalar Lisboa Norte, Lisbon, Portugal

Abstract

Cancer is a heterogeneous disease that may differ significantly in the population. Patients with similar histopathology may show different clinical outcomes posing a great challenge for both diagnosis and therapy decisions. Studying transcriptomic data is expected to bring novel insights about the molecular mechanisms that underlie cancer formation and progression, hence improving the accuracy with which patients are stratified and treated.

Gene expression data from two distinct datasets were analysed: colorectal cancer data (TCGA) and colon normal tissue data (GTEx). Preliminary analyses were performed in both datasets, namely Principal Component Analysis (PCA), survival and differential gene expression analysis. Results show differences in gene expression and survival outcome between colon cancer patients in different stages of the disease. In order to select variables based on their correlations between normal and tumoral samples we proposed Inverse-Twiner. For a certain variable in the network, the more different the correlation pattern between normal and tumoral datasets, less penalized it will be in the regularization term of Cox regression. These sparse Cox regression models will further allow to identify genes associated with high/low survival risk.

The estimation of reliable models allows to identify biomarkers associated with disease outcomes, which may have a strong impact in personalized healthcare.

Funding: Supported by national funds through Fundação para a Ciência e a Tecnologia (FCT) with reference UID/CEC/50021/2019, projects PTDC/EMS-SIS/0642/2014 and PTDC/CCI-CIF/29877/2017.
Computational analyses of the influence of evolutionary processes on the spatial genetic variation of Asian Modern Humans

C Branco*,1,2, M Arenas1,2

*caraujo@uvigo.es
1Department of Biochemistry, Genetics and Immunology, University of Vigo, Vigo, Spain.
2Biomedical Research Center (CINBIO), University of Vigo, 36310 Vigo, Spain.

Observed spatial genetic variation (hereafter genetic gradients) can provide insights about the past migrations of our species. Genetic gradients can be influenced by diverse evolutionary processes such as range expansion, range contraction or populations admixture. An interesting case of these influences can be found in the evolution of modern humans in Asia since this continent experienced diverse processes such as (i) a recent glacial period (GP) that induced range contractions of Paleolithic populations, (ii) admixture between Paleolithic and Neolithic populations and (iii) migration through long-distance dispersal (LDD). Studies based on real data revealed an Asian genetic gradient with an east-west (E-W) orientation but the causes for this orientation still remain unclear. Here, we performed spatially-explicit computer simulations of genetic data, coupled with principal component analyses (PCA) that provide the genetic gradients from the simulated data, to study the influence of the cited evolutionary processes on the Asian genetic gradient. In particular, we analyzed which of those processes produce the most realistic genetic gradient. We found that the observed E-W genetic gradient can be obtained in (1) pure Paleolithic populations that experienced the range contraction induced by GP; (2) admixed populations where the Paleolithic populations suffered the GP or the Neolithic populations expanded from the Middle East and East Asia. Additionally, simulations under LDD increased the noise of the estimated genetic gradients. Altogether we conclude that the glacial period, together with Neolithic expansions, have driven current Asian genetic gradients.
**Virulence and antibiotic resistance plasticity of *Arcobacter butzleri*: insights on the genomic diversity of an emerging human pathogen**

Joana Isidro¹, Susana Ferreira², Miguel Pinto¹, Fernanda Domingues², Mónica Oleastro³, João Paulo Gomes¹, Vítor Borges¹

¹Bioinformatics Unit, National Institute of Health Doutor Ricardo Jorge, Lisbon, Portugal
²CICS-UBI-Centro de Investigação em Ciências da Saúde, Universidade da Beira Interior, Covilhã, Portugal
³National Reference Laboratory for Gastrointestinal Infections, National Institute of Health Doutor Ricardo Jorge, Lisbon, Portugal

*Arcobacter butzleri* is a foodborne emerging human pathogen, frequently displaying a multidrug resistant character¹⁻³. Still, the lack of comprehensive genome-scale comparative analysis has limited our knowledge on *A. butzleri* diversification and pathogenicity. Here, we performed a deep genome analysis of *A. butzleri* focused on decoding its core- and pan-genome diversity and specific genetic traits underlying its pathogenic potential and diverse ecology. *A. butzleri* (genome size 2.07–2.58 Mbp) revealed a large open pan-genome with 7474 genes (about 50% being singletons) and a small but diverse core-genome with 1165 genes. It presents a plastic virulome (including newly identified determinants), marked by the differential presence of multiple adaptation-related virulence factors, such as the urease cluster *ureD(AB)CEFG* (phenotypically confirmed), the hypervariable hemagglutinin-encoding *hecA*, a type I secretion system (T1SS) harboring another agglutinin and a novel VirB/D4 T4SS likely linked to interbacterial competition and cytotoxicity. In addition, *A. butzleri* harbors a large repertoire of efflux pumps (EPs) and other antibiotic resistant determinants. We unprecedentedly describe a genetic mechanism of *A. butzleri* macrolides resistance (inactivation of a TetR repressor likely regulating an EP). Fluoroquinolones resistance correlated with Thr-85-Ile in GyrA and ampicillin resistance was linked to an OXA-15-like β-lactamase. Remarkably, by decoding the polymorphism pattern of the main antigen *PorA*, we show that *A. butzleri* is able to exchange *porA* as a whole and/or hypervariable epitope-encoding regions separately, leading to a multitude of chimeric PorA presentations that can impact pathogen-host interaction during infection. Ultimately, our unprecedented screening of short sequence repeats indicates that phase variation likely modulates *A. butzleri* key adaptive functions.

In summary, this study constitutes a turning point on *A. butzleri* comparative genomics revealing that this human gastrointestinal pathogen is equipped with vast and diverse virulence and antibiotic resistance arsenals that open a multitude of phenotypic fingerprints for environmental/host adaptation and pathogenicity.
TAD-GConTool and CNV-ConTool to assist prediction of phenotypic outcome of chromosomal rearrangements

Fino J, David D

Departamento de Genética Humana, Instituto Nacional de Saúde Dr. Ricardo Jorge, INSA IP. Lisbon, Portugal;

With the advance of genome sequencing technologies, it is currently possible to identify a large number of chromosomal or genomic structural variants in a single individual. Therefore, the validation and manual assessment of structural variants clinical significance becomes unpractical and time consuming when performed with previous methodologies.

In order to assist the validation process, we developed two clinically inspired bioinformatics tools - TADGConTool and CNV-ConTool. They were developed in python with a Common Gateway Interface that allows easy and user-friendly access through any standards compliant web browser (available at: http://dgrctools.insa.min-saude.pt/).

TAD-GConTool collects genomic information of breakpoint regions, using topological associated domains (TADs) as reference. It then accesses public databases to retrieve elements found inside TADs, and the associated clinical phenotypes, highlighting those causing dominant disorders.

CNV-ConTool searches for overlaps between patient-specific breakpoints and CNVs, and those reported in several public databases.

These tools were already successfully applied to about 40 cases studied under the project “Next-gen cytogenetics enters clinical care and annotates the human genome” (HMSP-ICT/0016/2013) and are now being made available to the broader scientific community. These tools allowed a faster and more informed evaluation of the genomic structural variants, helping select potential pathogenic variants, either by identifying phenotype-associated genes, or by overlapping deletions and duplications with already described benign or pathogenic CNVs.

As genome sequencing is becoming more and more a routine method for identification of chromosomal and genomic structural variants, such clinically oriented bioinformatics tools are crucial and represent the first level of analysis toward personalized genomic medicine.

This research was supported by national funds through FCT - Fundação para a Ciência e a Tecnologia, Research Grant HMSP-ICT/0016/2013.
Exploring One-Shot Learning for improving Drug-Discovery

Torres, Luis H. Magalhães¹, Monteiro, Nelson R. C.¹,², José L. Oliveira², Ribeiro, Bernardete¹, Joel P. Arrais¹

¹Department of Informatics Engineering (DEI), Center for Informatics and Systems of the University of Coimbra (CISUC), University of Coimbra, Coimbra, Portugal
²Department of Electronics, Telecommunications and Informatics, IEETA, University of Aveiro, 3810-193 Aveiro, Portugal

The application of deep neural networks in drug-discovery is mainly due to their enormous potential to significantly increase the predictive power when inferring the properties and activities of small-molecules. One major requirement to ensure the validity of the obtained neural networks models is the need for a large number of training examples per class. This invalidates the use of instances whose classes were not considered in the training phase or in data where the number of classes is high and oscillates dynamically. Unfortunately, this is a common scenario in drug-discovery, where the lead-optimization step is, inherently, a low-data problem, which makes it difficult to find potential analogous molecules with the desired therapeutic activity.

The main objective of this work is to optimize the discovery of drug analogues, with increased therapeutic activity, for the same pharmacological target, based on a reduced set of candidate drugs. We propose the use of a Siamese neural network architecture for one-shot classification, based on Convolutional Neural Networks (CNNs), that learns from a similarity score between two input molecules according to a given similarity function.

Using a one-shot learning strategy, we only need one instance per class for the network's training and a small amount of data and computational resources to build an accurate model. The preliminary results of this study showed that a one-shot learning strategy achieved consistent results given the low data available.

Acknowledgments

This research has been funded by the Portuguese Research Agency FCT, through D4 - Deep Drug Discovery and Deployment (CENTRO-01-0145-FEDER-029266)
Translational bioinformatics on daily diagnostics

Monteiro Mariana B.\textsuperscript{1} and Panadero Joaquín\textsuperscript{2}

\textsuperscript{1}BioBam Bioinformatics S.L, Valencia, Spain. \textsuperscript{2}Universidad de Valencia, Valencia, Spain.

Prior to the rapid development of molecular techniques for clinical diagnostics, hematologic and cancer diseases were evaluated according to the traditional morphology under microscope techniques. With the emergence of novel molecular diagnosis technologies (microarrays, RealTime qPCR and Next Generation Sequencing) bioinformatics play a new role improving human health through data mining analysis. Furthermore, translational bioinformatics helps on providing accurate and rapid genetic diagnostic leading to personal therapy treatments throughout effective pipelines for SNP calling, detection of chromosome and gene translocations, copy number variations (CNVs) and metagenomics.
Identification of genetic markers for traits of economic importance in cork oak using high-throughput SNP genotyping

Marta A. Antunes¹, Bruna Mendes¹, Hugo Magalhães¹, Ana Usiê¹,², Teresa Sampaio³, Fernanda Simões⁴, Helena Almeida³, António Marcos Ramos¹,²

¹ Centro de Biotecnologia Agrícola e Agro-alimentar do Alentejo (CEBAL)/Instituto Politécnico de Beja (IPBeja), Beja 7801-908, Portugal.
² MED – Mediterranean Institute for Agriculture, Environment and Development, CEBAL — Centro de Biotecnologia Agrícola e Agro-Alimentar do Alentejo, 7801-908 Beja, Portugal.
³ Instituto Superior de Agronomia, Universidade de Lisboa, Tapada da Ajuda, 1349-017 Lisboa, Portugal.
⁴ INIAV - Instituto Nacional de Investigação Agrária e Veterinária, Oeiras 2780-157, Portugal.

Cork oak (Quercus suber) is an evergreen tree of the Fagaceae family, and one of the main Portuguese forestry species. Besides its crucial environmental role, this tree also has a significant economic importance in the Mediterranean region, because of its ability to produce cork, a special thick bark used in the wine industry and in other industrial applications. In this study we report, for the first time, a set of genetic markers associated with cork quality and cork oak borer (Coroebus undatus) attacks. Cork quality was evaluated by measuring its thickness at different time points, along with a traditional industry quality scoring system. A total of 16, 8, and 4 SNPs were significantly associated with cork thickness before cooking, after first cooking, and after second cooking, respectively. Moreover, three SNPs displayed a significant effect on cork quality score. Furthermore, a total of five SNPs were significantly associated with the incidence of flathead oak borer. These results represent the first set of genetic markers identified for these important cork oak traits. In the future, marker assisted selection schemes may be potentially applied in cork oak, using information derived from these genetic markers.
Development of an integrated metabolic and transcriptional regulatory model for *Saccharomyces cerevisiae*

Paul Drescher, Rafael Carreira, Paulo Vilaça, Isabel Rocha

*Saccharomyces cerevisiae* (*S. cerevisiae*) is commonly used as a cell factory for research and industrial applications. The development of optimization strategies for *S. cerevisiae* was extensively driven by the establishment of genome-scale metabolic models. Additionally, to aid metabolic modeling, multiple attempts to infer transcriptional regulatory networks from expression data have been made. However, these data-based regulatory networks may show limited applicability. Therefore, we propose a general, knowledge-based approach involving a pipeline to evaluate regulatory interactions of transcription factors (TFs) and target genes from the YEASTRACT database, and form a regulatory network from these filtered datasets. So far, we filtered regulatory interactions based on two criteria: the existence of direct binding evidence of the TFs and their target genes, and the consensus over the regulatory effects identified among multiple studies that investigated each interaction, respectively. From 230 TFs and over 6000 genes contained in YEASTRACT, we obtained a regulatory network of 69 TFs and 1187 target genes with 1813 regulatory interactions, including 346 metabolic genes with 622 regulatory interactions. These interactions cover the majority of genes in the central carbon metabolism. Next, these regulatory interactions will be evaluated by identifying the network’s attractor states and simulating their effect on the metabolic system via steady-state regulatory flux balance analysis. Finally, this integrated metabolic and regulatory model may be used to identify efficient optimization strategies for *S. cerevisiae*.
Genomic characterization of emerging penicillin non-susceptibility among group B streptococci

Rafael Mamede¹, José Melo-Cristino¹, Mário Ramirez¹, Elisabete Martins¹

¹Instituto de Microbiologia, Instituto de Medicina Molecular, Faculdade de Medicina, Universidade de Lisboa, Lisbon, Portugal

Group B streptococci (GBS) have been considered uniformly susceptible to penicillin. In spite of this, the identification of penicillin-non-susceptible GBS (PRGBS) with mutations in the pbp genes has been occasionally but increasingly reported in Asia and North America, but was only recently reported in a single isolate in Europe.

Out of 252 GBS isolates recovered from vaginorectal swabs of pregnant Portuguese women, three showed decreased penicillin susceptibility (MIC>0.12 μg/ml). We used a High Throughput Sequencing (HTS) approach to characterize and compare the three PRGBS isolates, and de novo assembly was performed with the INNUca pipeline. Sequence Type (ST) determination, genome annotation and variant calling were carried out with MLST, Prokka and Snippy, respectively. A core-genome MultiLocus Sequence Typing analysis was performed with the chewBBACA suite and the allelic profiles were imported to PHYLOViZ Online for visualization. Additionally, genomic regions corresponding to capsular serotype determining genes, surface-associated structures and pbp genes were identified with BLAST.

The genomic analysis revealed that all PRGBS isolates contained the type III capsular polysaccharide locus, carried the rib surface protein gene, pilus islands 1 and 2b, and were represented by ST109, a single-locus variant of ST17, founder of the hypervirulent CC17 clone. Comparison of the PBPs with those of the susceptible strain 2603V/R (NC_004116.1) revealed, in all isolates, the presence of 4, 3, 1 and 2 amino acid substitutions in pbp1a, pbp2a (including a G398A substitution previously found among PRGBS isolates in Japan), pbp2b and pbp2x, respectively. The core-genome MLST comparison of the 3 isolates with a reference genome (COH1, ST17) and a penicillin susceptible representative of ST17 in the collection revealed that the ST109 isolates are more similar to each other (between 46 and 86 allelic differences) than to the ST17 isolate, which is closer (94 allelic differences) to the COH1 reference.

The possible emergence of penicillin resistance in GBS, particularly among the most prevalent and invasive clones, such as the hypervirulent CC17 clone, may have significant implications for the prevention and management of GBS disease and a thorough analysis of identified cases is of crucial importance.
Acetylation’s role in tau structure, electrostatics and interactions: molecular dynamics studies

Tiago Ferreira¹, Tarsila G. Castro¹,², Florentina-Daniela Munteanu² and Artur Cavaco-Paulo¹,²

¹Centre of Biological Engineering, University of Minho, Campus de Gualtar, 4710-057, Braga, Portugal ²Aurel Vlaicu University of Arad, Str. Elena Dragoi 2-4, RO-310330 Arad, Romania;

Tau is a microtubule associated protein which stabilizes and promotes the assembly of microtubules in neurons.¹ In its functional form it presents minor modifications, such as phosphorylation, but a large variety of post-translational modifications are also possible.²

Tau post-translational modifications are directly related with tau mal-functioning, aggregation and subsequent tauopathies. Investigation on tau hyperphosphorylation has dominated in the past years, since its known role in Alzheimer's disease onset and progression. However, acetylation has gained attention, as this process is also responsible for tau pathological aggregation.³⁻⁴

Acetylation is a modification that occurs in lysine amino acids, adding an acyl group to the side chain NH moiety.⁵ This process changes the charge of this residue, making it neutral and consequently modifying the electrostatics of the whole protein. The presence of charged groups and electrostatic interactions are the major contributors for a final protein fold. The absence of these charges, via acetylation, will contribute to significant changes in tau’s structure and interactions.

Molecular Dynamics (MD) simulations take advantage of precise simulation algorithms and present themselves as a robust way to understand biomolecules’ behavior, conformational preferences and interactions, even for intrinsically disordered proteins such as tau. Through this technique we are following the acetylation impact on the tau structure and its way of binding to the microtubule. In addition, it is intended to unveil the relationship between acetylation and aggregation, which results in tau associated diseases.

In the past year, Castro and co-workers shed light on tau properties by predicting its 3D structure and disclosing the interaction pattern with microtubules and the ions from the intracellular fluid.¹ The present work took this input information to generate acetylated analogs and follow the impact of this modification on tau’s dynamic behavior.

Keywords: tau, molecular dynamics simulation, acetylation, Alzheimer’s disease.
Deep Reinforcement Learning Framework for Drug Design with Optimized Drug-Like Properties

Pereira, T.O\textsuperscript{1}, Abbasi Maryam\textsuperscript{1}, Arrais J.P.\textsuperscript{1}, Ribeiro B.\textsuperscript{1}

\textsuperscript{1}Department of Informatics Engineering (DEI), Center for Informatics and Systems of the University of Coimbra (CISUC), University of Coimbra, Coimbra, Portugal

In de novo drug design, computational strategies have been used to generate new molecules with bespoke properties that have a good affinity towards the desired biological purpose.

In this work, we explore the use of Reinforcement Learning (RL) strategies for improving the design of de novo drug-like compounds. Basically, RL is a framework where an agent interacts with the environment through its actions and receives a numerical reward. Its aim is to maximize a long-term objective.

Firstly, the generative model is built using Recurrent Neural Networks (RNN). We perform the training process of this model with a set of SMILES sufficiently representative of the building rules of valid SMILES. We use a Long Short-Term Memory (LSTM) architecture with two layers that are trained to learn how to generate valid molecules using its ability to predict future actions based on the previous information. The LSTM layers are followed by a dense layer and a neuron unit with a SoftMax activation function to normalize the output into a probability distribution.

Afterwards, we implemented the evaluator using two LSTM layers to estimate the desired property of each generated structure. Conversely, the output was the physical, chemical or biological SMILE property and, as such, we used a dense layer with just one unit for this purpose.

Subsequently, the policy gradient method is applied to make the model produce fine-tuned molecules. To do that, it is necessary to maximize the objective function by updating the generator’s weights according to the policy gradient algorithm, increasing the probability of molecules with higher rewards and avoiding generating those molecules that provide lower rewards.

The performed experiments demonstrate the efficiency of the proposed strategy in a single task regime where each endpoint of interest is independently optimized. For instance, regarding the coefficient partition - which is the measure of how the candidate drug can be absorbed through the cell membrane - we managed to increase the percentage of generated drugs with admissible solubility, according to Lipinski’s rule.

Nevertheless, this approach can be expanded simultaneously to allow multi-objective optimization of several drug-like properties, which is the need for drug discovery.

Acknowledgements: This research has been funded by the Portuguese Research Agency FCT, through D4 - Deep Drug Discovery and Deployment (CENTRO-01-0145-FEDER- 029266).
BioData.pt User Support: Services and Software

Neves, D¹, Costa, Henrique¹

¹ Instituto Gulbenkian de Ciência

Modern biological and bioinformatics research is often dependant on analyzing large amounts of data, which requires a specific skillset. BioData’s user support seeks to empower users, whether they come from research or industry, by developing software which lets the user analyze large data through user-friendly applications. In addition, we provide one-day Crash Courses, to introduce participants to concepts and software that they can use for their work.
A reconciled version of the cork oak tree genome-scale metabolic model

Miguel Silva¹, Diogo Lima¹, Emanuel Cunha¹, Oscar Dias¹, Inês Chaves², Célia Miguel², Miguel Rocha¹ and Isabel Rocha²

¹ BIOSYSTEMS, Centre of Biological Engineering, University of Minho, Campus de Gualtar 4710-057 Braga Portugal
² ITQB NOVA, Av. da República, 2780-157 Oeiras, Portugal

Quercus suber, commonly known as cork oak tree, is an evergreen tree which produces a thick bark (also known as cork) with multiple (a)biotic stress resistance properties (1). Due to cork’s natural characteristics, such as the low weight, excellent insulation and low permeability, the cellular structure has a significant economic value as it has multiple applications. For instance, it can be used as a wine bottle sealant and insulation boards (2,3). Additionally, cork is harvested periodically throughout the tree’s lifetime (4). Nevertheless, the cork’s quality can only be properly assessed after 40 years of tree growth, which makes the identification of metabolic traits, associated to high-quality cork, of the utmost importance (5).

Genome-Scale Metabolic (GSM) models comprise both genomic and metabolic information and can predict the phenotypic behavior of an organism when subjected to distinct environmental conditions (6). Therefore, a reconstructed GSM model of the cork oak tree can point to metabolic properties related to cork quality. Additionally, in silico metabolic engineering strategies could lead to the development of metabolically enhanced trees.

The current Quercus suber leaf model, reconstructed within merlin (7), contains 3126 reactions, 2648 metabolites, 7258 genes and was subjected to extensive manual curation, while the biomass and energy requirements were revamped. In silico simulations, using Flux Balance Analysis (8), accurately predict the phenotypic behavior of the leaf cell when exposed to phototrophic and heterotrophic conditions.