Pangenome wide prediction of gene function

Prof. Marco Galardini
Hannover Medical School / Twincore / HZI
Twincore, Room 2170
Feodor-Lynen-Straße 7
30625 Hannover
Email: marco.galardini@twincore.de

The "Systems Biology of Microbial Communities" lab at Twincore/MHH is looking for a computational biology PhD student to undertake research in the area of microbiology and machine learning, with the possibility of also acquiring molecular biology laboratory experience.

Marco Galardini is a computational biologist with a taste for microbiology. He received his PhD in bioinformatics and microbiology at the University of Florence (with Prof. Bazzicalupo), studying the complex genome of the rhizobial species *S. meliloti*. He then did a postdoc at EMBL-EBI in Cambridge (with Dr. Beltrao) studying the genotype-to-phenotype relationship in an *E. coli* strain collection, as well as the variability in gene essentiality profiles across *S. cerevisiae* strains. Lastly, he has spent the last year and a half learning about high-throughput laboratory evolution in the lab of Prof. Khalil at Boston University. He has started his position as Associate (W2) Professor in Systems Biology of Microbial Communities at MHH/Twincore/HZI in October 2020.

Research Interests and Achievements

Prof. Galardini has been fascinated by microbiology since his undergraduate studies, and has spent his research career trying to understand the extraordinary diversity of the bacterial kingdom. The high plasticity of bacterial genomes (collectively termed "pangenomes") and their relationship with changes in phenotypes forms the core of his research. In particular, he is interested in understanding how genotype changes within a bacterial species translates to changes in phenotypes and how those affect important biological processes such as infectious diseases. The availability of cheap whole genome sequencing and high-throughput molecular and phenotyping platforms has made these kinds of studies now possible at scale. In fact, Prof. Galardini has built an *E. coli* strain collection to study various aspects of the genotype-to-phenotype relationship, such as growth in stressful conditions (Galardini *et al.*, eLife 2017) as well in pathogenicity (Galardini *et al.*, 2019 bioRxiv) and evolution of antimicrobial resistance (unpublished). Future plans for this strain collection include extending the genotype-to-phenotype maps to include other molecular processes (transcription, translation, metabolism, …), studying the influence of the genetic background on variant effects (as in Galardini *et al.*, 2019 MSB) and evolution.

Funding

The lab has funding for the next 5 years from the RESIST excellence cluster, with the potential for an additional 5-year extension pending a successful review.

Selected Publications

- **Galardini, M.**, Koumoutsi, A., Herrera, C. M., Cordero Varela, J. A., Telzerow, A., Wagih O., Wartel M., Clermont O., Denamur E., Typas, A., Beltrao, P. (2017). **Phenotype inference in an *Escherichia coli* strain panel.** *eLife*, *6*.

- **Galardini, M.**, Busby, B. P., Vieitez, C., Dunham, A. S., Typas, A., & Beltrao, P. (2019). **The impact of the genetic background on gene deletion phenotypes in *Saccharomyces cerevisiae*.** Molecular Systems Biology, 15(12).

- **Galardini, M.**, Clermont, O., Baron, A., Busby, B., Dion, S., Schubert, S., ... & Denamur, E. (2019). **Major role of the high-pathogenicity island (HPI) in the intrinsic extra-intestinal virulence of Escherichia coli revealed by a genome-wide association study.** *BioRxiv*, 712034.

- Wagih, O., **Galardini, M.**, Busby, B. P., Memon, D., Typas, A., & Beltrao, P. (2018). **A resource of variant effect predictions of single nucleotide variants in model organisms.** *Molecular systems biology*, *14*(12), e8430.

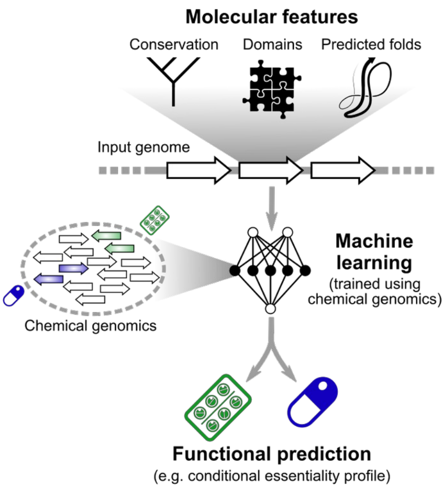
- **Galardini, M.**, Brilli, M., Spini, G., Rossi, M., Roncaglia, B., Bani, A., ... & Pini, F. (2015). **Evolution of intra-specific regulatory networks in a multipartite bacterial genome.** *PLoS Comput Biol*, *11*(9), e1004478.

Background

Prof. Galardini's previous work in genotype-to-phenotype associations and predictions have demonstrated the possibility to infer the impact of genetic variants on phenotype. In particular he has shown how a relatively simple mechanistic model based on the impact of non-synonymous variants can predict growth defects across multiple conditions (Galardini *et al.*, 2017 eLife). There is however one big caveat in such models: the influence of the so-called accessory genome – that is, those genes that are not present in the reference strain – cannot be taken into account. As the functions of those genes is largely unknown, there is a pressing need to understand how the accessory genome influences phenotype. Since molecular laboratory techniques to investigate gene function do not scale well with the tens of thousand of accessory genes belonging to a species' pangenome, computational methods such as machine learning could be used to predict gene function *in silico*.

Project description

Computational approaches such as machine learning trained on the wealth of data available for model organisms and using features extracted from nucleotide sequences as input could be used to improve the current function prediction methods. The project then would involve developing a predictor of gene function for bacterial accessory genes, using a machine learning model trained on available chemical genomics data. The input of the model would be the genome sequence of an isolate, while the output would be a probability that each input gene is essential for growth in a series of conditions. As an example, a model for gene function in the *E. coli* species could be constructed by using the chemical genomics data that is available for the model strain K-12 as training; the genome sequence of another *E. coli* strain with unannotated genes could then be used as input to this model to generate a functional prediction for each gene. As chemical genomics data is now available for many bacterial species, it would be possible to generate a model specific to each species, which could then be applied to all available and future genome sequences of strains belonging to the species.



A large number of molecular features can be extracted from genome sequences and used as the input for functional prediction. Homology to all known protein coding genes can be computed at scale thanks to curated databases and fast algorithms; the degree of conservation of a gene across bacterial species is by itself a strong predictor of a biologically-relevant function. The gene sequence can also be further segmented to look for conserved domains, which are also a powerful indicator for a more specific function. Recent advancements in protein structure prediction from co-evolution profiles offer an additional, finer grained source of information for functional prediction; certain protein folds can in fact be conductive of the function of a gene, even if no sequence homology to known proteins is present. Generating protein structure predictions and using them for gene function prediction has not yet been attempted at the genomic scale, and it could improve significantly what can currently be functionally predicted from sequence alone. Genome organization is an additional feature that can be easily computed and used as input for the model, leveraging the structural conservation of bacterial genomes, as well as gene co-occurrence patterns.

Traditional machine learning models such as random forests will be used to train the predictor, aided by powerful deep learning models such as auto-encoders, which are able to generate higher-order representations of the input data by a process of dimensionality reduction, thus reducing the noise of the data fed into the model. The predictions will be validated through standard cross-validation and through the generation of new knock-out libraries in a few strains, either in the lab or through an existing collaboration with Nassos Typas (EMBL). The final predictors for each species will be released as a freely available web-service, to the benefit of the bacterial community.