SysbioBig

Thesis Proposal

Università degli Studi di Padova

Dipartimento di Ingegneria dell'Informazione

SysBioBiG

Bivariate statistic extension to account for multiple covariates in genotype-phenotype association

Objective:
Develop a methodology to integrate multiple covariates (e.g., age, sex, environmental factors) into the existing bivariate entropy-based statistic used in ABACUS, assessing their impact on genotype-phenotype associations and evaluating whether accounting for these factors improves statistical power while reducing false positives in rare variant detection.

Skill requirements:

- Familiarity with **C / C++**
- Proficiency in **R** and/or **Python** for statistical modeling

Preferred background:

- Computer engineering/science
- Bioengineering

$$S_2(snp_A, snp_B) = \sum_{g=1}^{9} \frac{H_0 - H_g}{H_0} \cdot F_g$$

$$H_g = -\sum_c \frac{f_{cg}}{\sum_c f_{cg}} \log_2 \left(\frac{f_{cg}}{\sum_c f_{cg}} \right)$$

CONTINGENCY TABLE FOR GENOTYPE CONFIGURATION OF A SNP PAIR.

	AABB	AABb	AAbb	AaBB	AaBb	Aabb	aaBB	aaBb	aabb
C_1	f_{11}	f_{12}	f_{13}	f_{14}	f_{15}	f_{16}	f_{17}	f_{18}	f_{19}
C_2	f_{21}	f_{22}	f_{23}	f_{24}	f_{25}	f_{26}	f_{27}	f_{28}	f_{29}

How can covariates influence the genotype entropy?

References/useful resources:

Di Camillo B, Sambo F, Toffolo G, Cobelli C. ABACUS: an entropy-based cumulative bivariate statistic robust to rare variants and different direction of genotype effect. Bioinformatics. 2014 Feb 1;30(3):384-91. 10.1093/bioinformatics/btt697

Uffelmann E, Huang QQ, Munung NS, De Vries J, Okada Y, Martin AR, Martin HC, Lappalainen T, Posthuma D. Genome-wide association studies. Nature Reviews Methods Primers. 2021 Aug 26;1(1):59. 10.1038/s43586-021-00056-9

Contacts: barbara.dicamillo@unipd.it
mikele.milia@phd.unipd.it

SysBiOBiG

Synthetic generation of phenotype-genotype contingency matrices for null model optimization

Objective:
Development of a framework to use essential SNP parameters to generate synthetic genotype-phenotype (case/control) contingency tables, reducing the need for direct computation from raw genetic data to construct a null model for performing SNP-phenotype association. Then evaluate it against existing ABACUS framework to assess its impact on computational efficiency, memory usage, and accuracy of null model estimation for large-scale genomic data.

Skill requirements:

- Familiarity with **C / C++**
- Proficiency in **R** and/or **Python** for statistical modeling

Preferred background:

- Computer engineering/science
- Bioengineering

Allele Frequencies: $p, q = 1 - p$

Genotype Frequencies: $P(AA)=p^2, P(Aa)=2pq, P(aa)=q^2$

Phenotype distribution (case / control ratios)

Linkage disequilibrium: D', r^2

CONTINGENCY TABLE FOR GENOTYPE CONFIGURATION OF A SNP PAIR.

	AABB	AABb	AAbb	AaBB	AaBb	Aabb	aaBB	aaBb	aabb
C_1	f_{11}	f_{12}	f_{13}	f_{14}	f_{15}	f_{16}	f_{17}	f_{18}	f_{19}
C_2	f_{21}	f_{22}	f_{23}	f_{24}	f_{25}	f_{26}	f_{27}	f_{28}	f_{29}

How can we compute a null model using only few parameters? Is it possible to infer exact frequencies?

References/useful resources:

Di Camillo B, Sambo F, Toffolo G, Cobelli C. ABACUS: an entropy-based cumulative bivariate statistic robust to rare variants and different direction of genotype effect. Bioinformatics. 2014 Feb 1;30(3):384-91. 10.1093/bioinformatics/btt697

Uffelmann E, Huang QQ, Munung NS, De Vries J, Okada Y, Martin AR, Martin HC, Lappalainen T, Posthuma D. Genome-wide association studies. Nature Reviews Methods Primers. 2021 Aug 26;1(1):59. 10.1038/s43586-021-00056-9

Contacts: barbara.dicamillo@unipd.it
mikele.milia@phd.unipd.it

SysBiOBiG

Multi modal data integration of snRNA-seq and im-scRNA seq data

Objective:

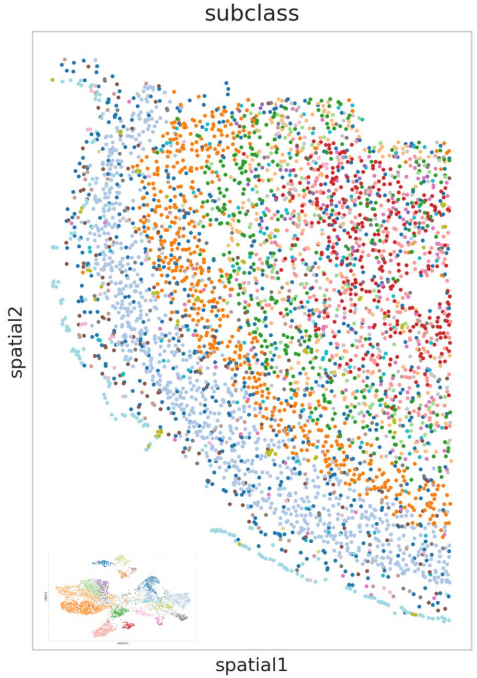
Adaptation of a loss function used to find the optimal map to redistribute the transcriptomic profile of snRNA seq data (10x V3) based on a reference given from im-scRNA seq data (MERFISH) of a Mouse MOp cortex dataset.

Skill requirements:

- Fundamentals of machine learning
- Basic skills in Python
- Attention to big data handling

Preferred background:

- Computer engineering/science
- Bioengineering



snRNA seq data (rectangular-ish matrix) with im-scRNA seq (very tall matrix)

$X_c \in \mathbb{R}^{C \times G}$ $X_s \in \mathbb{R}^{S \times G'}$

Predicting undetected transcripts in spatial domains by minimizing LOSS Φ(M)

$$\Phi\left(\tilde{M}\right) = KL\left(\vec{\mathrm{m}}, \vec{\mathrm{d}}\right) - \sum_{k}^{n_{genes}} cos_{sim}\left((M^T X_c)_{*,k}, X_{s_{*,k}}\right) - \sum_{j}^{n_{voxels}} cos_{sim}\left((M^T X_c)_{j,*}, X_{s_{j,*}}\right),$$

Obtain a pseudo spatial transcriptomics gene expression matrix

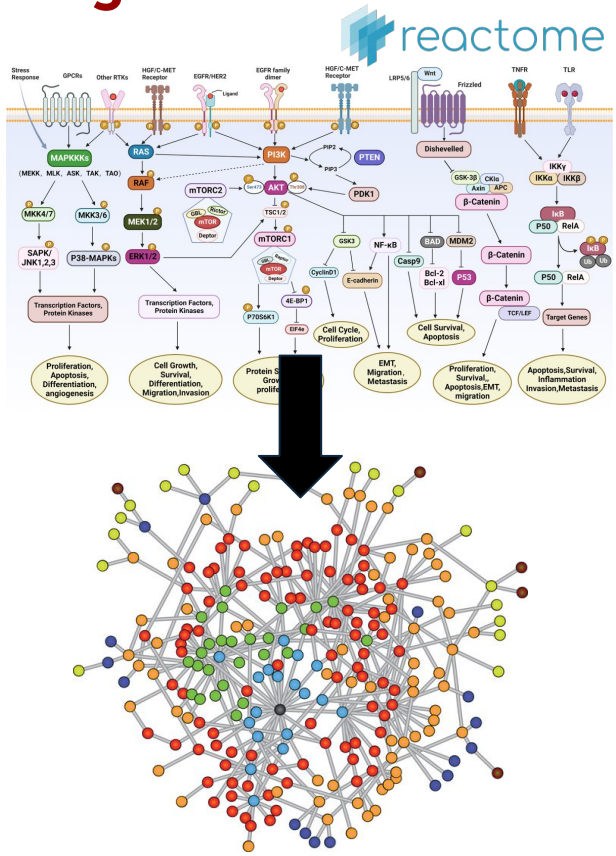
References/useful resources:

Biancalani, T., Scalia, G., Buffoni, L. *et al.* Deep learning and alignment of spatially resolved single-cell transcriptomes with Tangram. *Nat Methods* **18**, 1352–1362 (2021). https://doi.org/10.1038/s41592-021-01264-7

Li, B., Zhang, W., Guo, C. *et al.* Benchmarking spatial and single-cell transcriptomics integration methods for transcript distribution prediction and cell type deconvolution. *Nat Methods* **19**, 662–670 (2022). https://doi.org/10.1038/s41592-022-01480-9

Contacts: barbara.dicamillo@unipd.it
matteo.baldan.7@phd.unipd.it

Conversion of signaling pathway into gene-gene network



Objective:

Develop a biologically-driven methodology to convert information from Reactome DB into a gene-gene interaction network

- Analyze the structure and organization of Reactome data (JSON format)
- Define mathematical rules to represent biological entities (protein complexes, genes, and metabolites) as nodes and chemical reactions as edges while preserving network topology and biological information.

Skill requirements:

- Fundamentals of network theory
- Basic programming skills in R and/or Python

Preferred background:

- Computer engineering/science
- Bioengineering

References/useful resources:

Sales, G., Calura, E., Cavalieri, D. *et al.* graphite - a Bioconductor package to convert pathway topology to gene network. *BMC Bioinformatics* 13, 20 (2012). https://doi.org/10.1186/1471-2105-13-20

Contacts: barbara.dicamillo@unipd.it
giulia.cesaro@unipd.it

SysBiOBiG

Memory and time efficient representation and computation on sparse matrix

Objective:

Implement a R library designed to store and process large sparse matrices, leveraging C++ for efficient data structures and parallel computation.

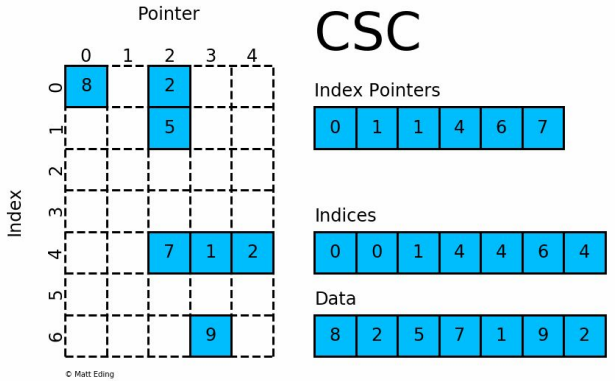
- Implement a C++ data structure for sparse matrix representation
- Design and optimize computationally efficient algorithms for sparse matrix operations (e.g. statistical tests).
- Integrate parallel programming using OpenMP (shared-memory)

Skill requirements:

- Proficiency in C++/C programming

Preferred background:

- Computer engineering/science



References/useful resources:

http://www.stat.yale.edu/~mjk56/temp/bigmemory-vignette.pdf

https://www.rcpp.org/

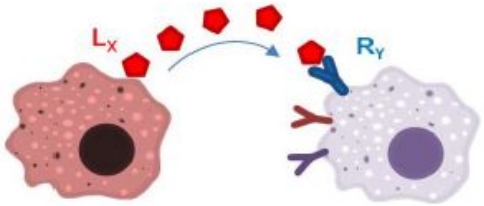
Contacts: barbara.dicamillo@unipd.it
giacomo.baruzzo@unipd.it
giulia.cesaro@unipd.it

Develop benchmarking framework for cell-cell communication inference tools

Objective:

Implement and develop a benchmarking framework for cell-cell communication inference tools using omics data

- Identify test scRNA-seq datasets for validation using complementary omics data or controlled environments
- Identify quantitative metrics for assessing the accuracy, robustness and computational efficiency
- Develop a computational framework to systematically compare existing tools



Skill requirements:

- Basic programming skills in R and/or Python
- Fundamentals of statistics

Preferred background:

- Computer engineering/science
- Bioengineering

References/useful resources:

Luo J, Deng M, Zhang X, Sun X. ESICCC as a systematic computational framework for evaluation, selection, and integration of cell-cell communication inference methods. Genome Res. https://www.genome.org/cgi/doi/10.1101/gr.278001.123.

Dimitrov, D., Türei, D., Garrido-Rodriguez, M. *et al.* Comparison of methods and resources for cell-cell communication inference from single-cell RNA-Seq data. *Nat Commun* 13, 3224 (2022). https://doi.org/10.1038/s41467-022-30755-0

Contacts: barbara.dicamillo@unipd.it
giulia.cesaro@unipd.it

SysBioBiG

Bacterial community simulation

Objective:

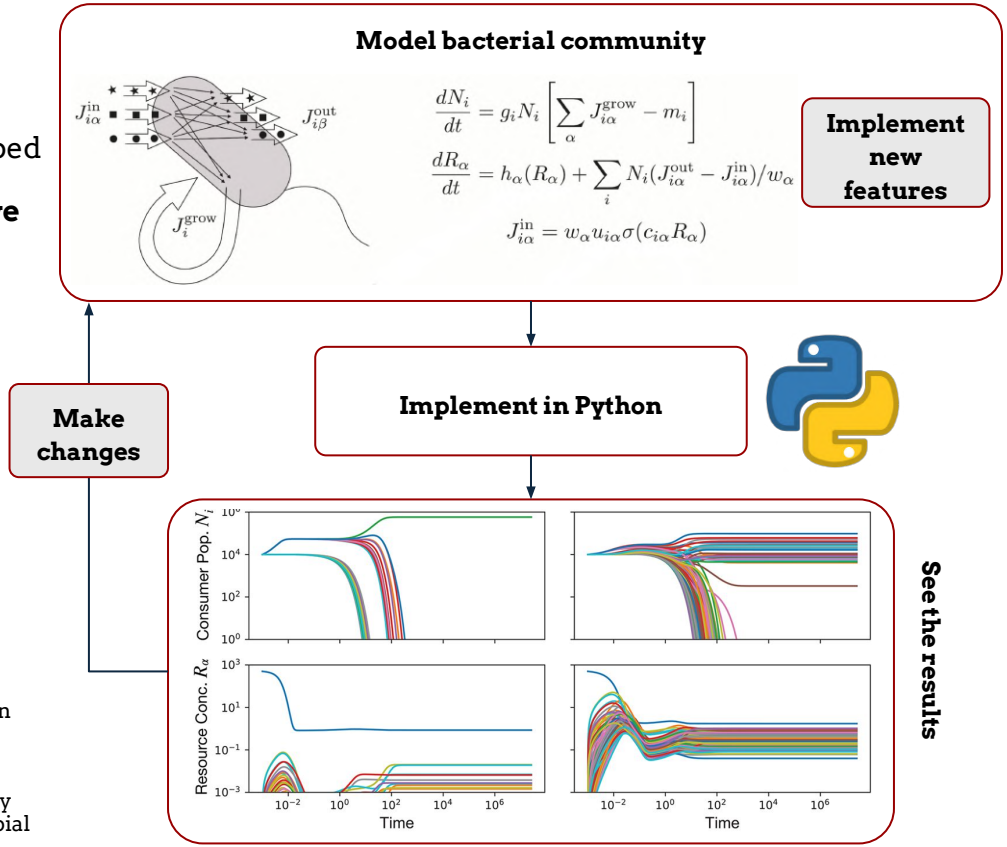
Improve and expand bacterial community simulator developed by our lab. The main goals are to **expand the simulator functionalities** (predation, bacterial families,...) and **improve numerical stability** and speed.

Skill requirements:

- Modelling
- Code implementation (Python)
- Code optimization

Preferred background:

- Computer engineering/science
- Bioengineering



References/useful resources:

Marsland R, Cui W, Goldford J, Mehta P (2020) The Community Simulator: A Python package for microbial ecology. PLOS ONE 15(3): e0230430. https://doi.org/10.1371/journal.pone.0230430

Marsland R III, Cui W, Goldford J, Sanchez A, Korolev K, et al. (2019) Available energy fluxes drive a transition in the diversity, stability, and functional structure of microbial communities. PLOS Computational Biology 15(2): e1006793. https://doi.org/10.1371/journal.pcbi.1006793

Contacts: barbara.dicamillo@unipd.it
piero.mariotto@phd.unipd.it

Analysis of sequencing data to evaluate the efficacy of Faecal Microbiota Transplant

Objective:

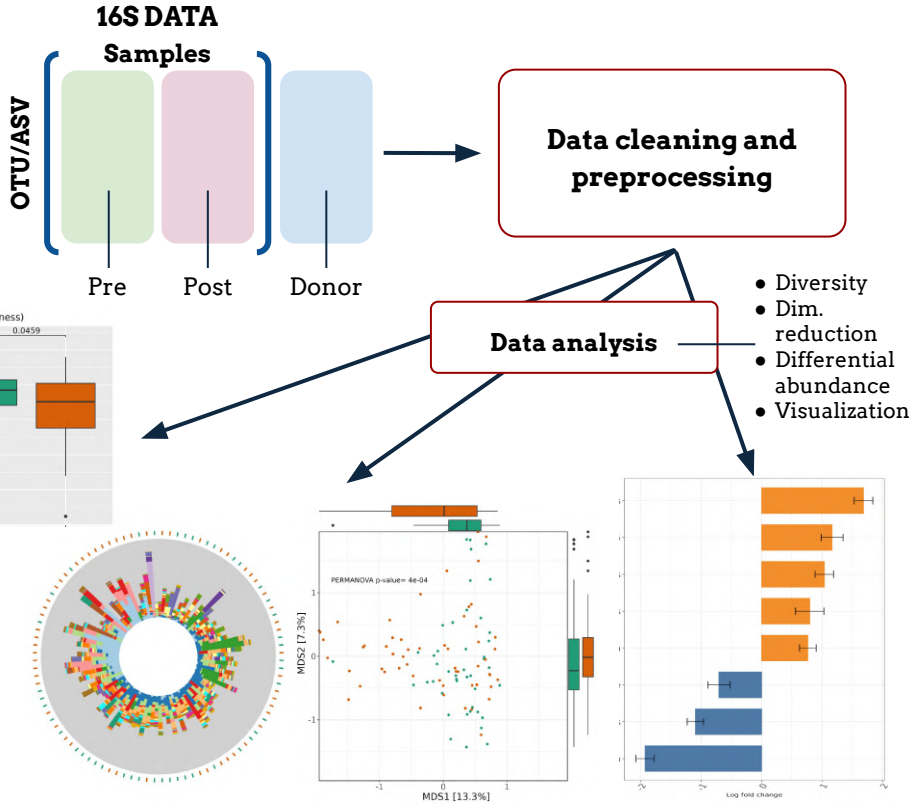
Analyse 16S sequencing data to evaluate the effect of FMT on the gut microbiota of dogs and cats. The main objective is to **validate FMT** as a treatment and to individuate covariates that influence the outcome.

Skill requirements:

- Data handling
- Data analysis (using mainly R)
- Statistical testing

Preferred background:

- Computer engineering/science
- Bioengineering



References/useful resources:

Yadegar A, Bar-Yoseph H, Monaghan TM, Pakpour S, Severino A, Kuijper EJ, Smits WK, Terveer EM, Neupane S, Nabavi-Rad A, Sadeghi J, Cammarota G, Ianiro G, Nap-Hill E, Leung D, Wong K, Kao D.2024.Fecal microbiota transplantation: current challenges and future landscapes. Clin Microbiol Rev 37:e00060-22. https://doi.org/10.1128/cmr.00060-22

Berlanda, M.; Innocente, G.; Simionati, B.; Di Camillo, B.; Facchin, S.; Giron, M.C.; Savarino, E.; Sebastiani, F.; Fiorio, F.; Patuzzi, I. Faecal Microbiome Transplantation as a Solution to Chronic Enteropathies in Dogs: A Case Study of Beneficial Microbial Evolution. *Animals* **2021**, *11*, 1433. https://doi.org/10.3390/ani11051433

Contacts: barbara.dicamillo@unipd.it
piero.mariotto@phd.unipd.it

Inference of microbial genomic data: a benchmark of existing tools

Objective:

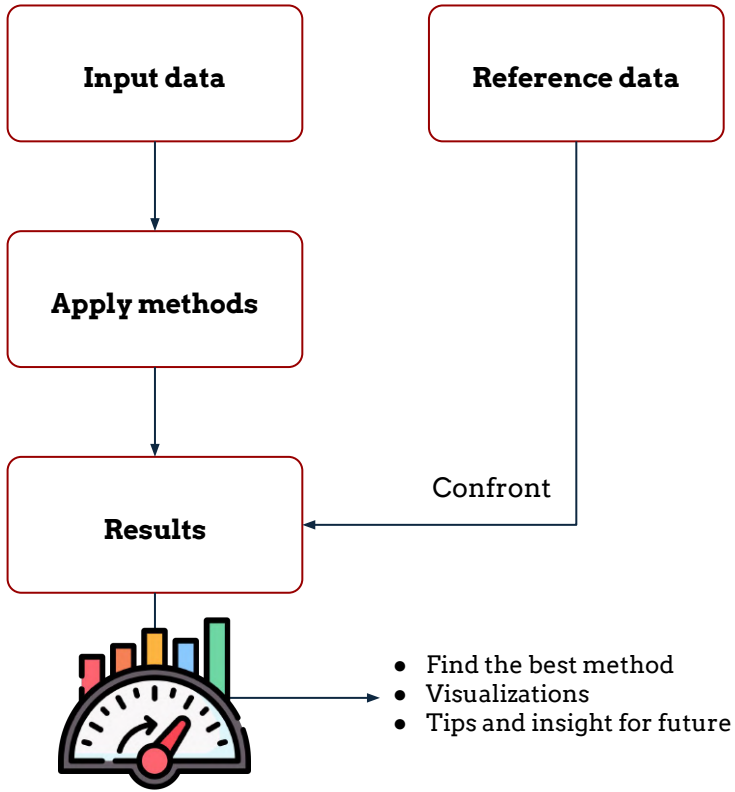
Sequencing of the whole genome of bacteria is costly and complex. For this reason, many tools have been developed to **infer genome sequences from 16S data**. Our objective is to perform a comprehensive **benchmark of existing methods**.

Skill requirements:

- Careful and systematic development of benchmark pipeline
- Attention to data handling
- Programming skills (Python & R)

Preferred background:

- Computer engineering/science
- Bioengineering



References/useful resources:

Sun, S., Jones, R.B. & Fodor, A.A. Inference-based accuracy of metagenome prediction tools varies across sample types and functional categories. *Microbiome* **8**, 46 (2020). https://doi.org/10.1186/s40168-020-00815-y

Christophe Djemiel, Pierre-Alain Maron, Sébastien Terrat, Samuel Dequiedt, Aurélien Cottin, Lionel Ranjard, Inferring microbiota functions from taxonomic genes: a review, *GigaScience*, Volume 11, 2022, giab090, https://doi.org/10.1093/gigascience/giab090

Contacts: barbara.dicamillo@unipd.it
piero.mariotto@phd.unipd.it

Benchmarking TFs activity inference methods

Objective:

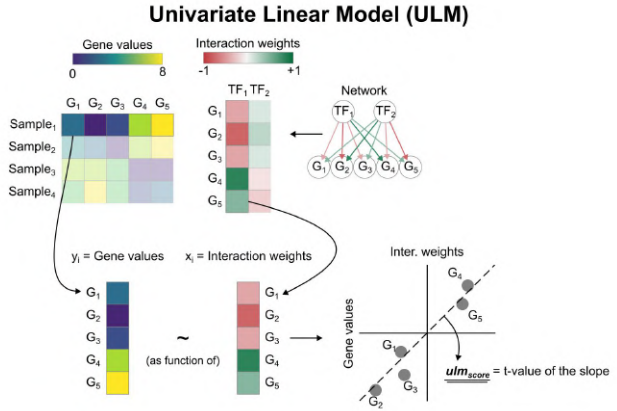
Single-cell RNA sequencing data allow to gain deep insights on the functional state of cells. At the state of the art different methods have been developed to infer the activity of **transcription factors** - TFs, regulators of gene expression - implementing different mathematical formulations. Our objective is to perform a comprehensive **benchmark** of existing methods, considering both the biological accuracy and the mathematical assumptions of the different approaches.

Skill requirements:

- Attention to data handling
- Programming skills (Python & R)
- Statistical background

Preferred background:

- Computer engineering/science
- Bioengineering



References/useful resources:

Mompel et al., decoupleR: ensemble of computational methods to infer biological activities from omics data, Bioinformatics Advances, Volume 2, Issue 1, 2022, vbac016, https://doi.org/10.1093/bioadv/vbac016

Müller-Dott et al., Expanding the coverage of regulons from high-confidence prior knowledge for accurate estimation of transcription factor activities bioRxiv 2023.03.30.534849; doi: https://doi.org/10.1101/2023.03.30.534849

Contacts: barbara.dicamillo@unipd.it
gaia.tussardi@phd.unipd.it

Development of ML-based models of disease progression using clinical data

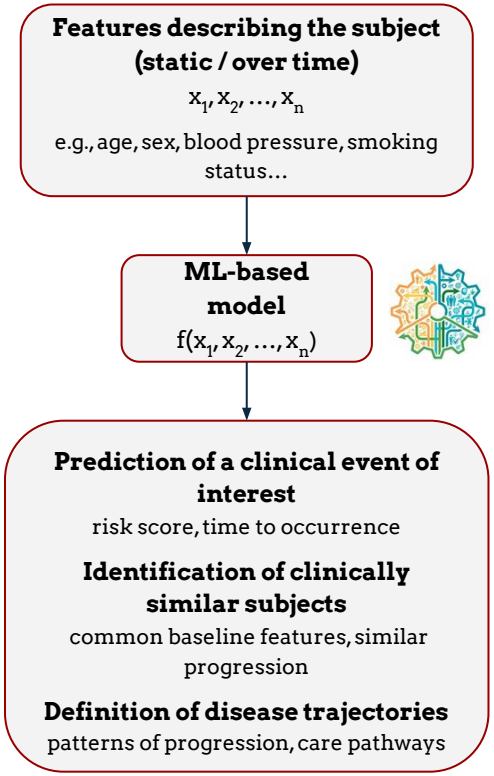
Objective:
Development of machine learning-based approaches for modelling disease progression and patient stratification, aiming to forecast disease trajectories and predict adverse events, with a focus on enhancing personalised healthcare strategies.

Hard/soft skills requirements:

- Fundamentals of machine learning and data analysis
- Basic programming skills in R and/or Python
- Attention to data quality and curation

Preferred background:

- Computer engineering/science
- Bioengineering
- Data science



References/useful resources:

Tavazzi et al, 2023: Artificial intelligence and statistical methods for stratification and prediction of progression in amyotrophic lateral sclerosis: A systematic review. 10.1016/j.artmed.2023.102588

Trescato et al, 2024: DYNAMITE: Integrating Archetypal Analysis and Process Mining for Interpretable Disease Progression Modelling 10.1109/JBHI.2024.3453602

Contacts: barbara.dicamillo@unipd.it
erica.tavazzi@unipd.it

Network Analysis of Patient Data: Community Detection and Temporal Evolution

Community Detection:

Objective:

Analyse patient networks using community detection algorithms, and study the evolution of network structures, including changes in centrality measures, adjacency matrices, and edge weight distributions.

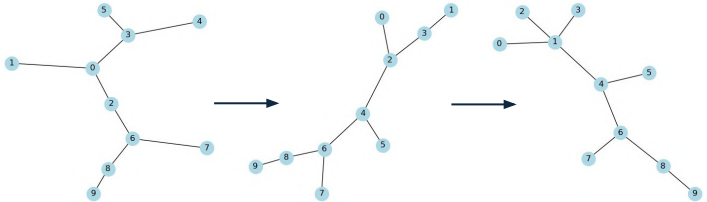
Hard/soft skills requirements:

- Fundamentals of network analysis and machine learning
- Basic programming skills in Python and/or R

Preferred background:

- Computer engineering/science
- Bioengineering
- Data science
- Physics

Temporal evolution:



References/useful resources:

Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, Etienne Lefebvre. Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment, 2008, P10008, pp.1-12. https://doi.org/10.1088/1742-5468/2008/10/P10008

https://networksciencebook.com/

Contacts: barbara.dicamillo@unipd.it
elena.marinello.1@phd.unipd.it