MENU: Home, Description, Members, Publications, Activities

**SNPs-PHE: Computational models for predicting SNPs-phenotype associations**

**STIC-AmSud Project (2017-2018)**

# En menu Home:

# [https://date.ac.lri.fr/img/sticamsud.png](http://www.sticamsud.org/)

# STIC-AmSud Program

The Regional Program STIC-AmSud is an initiative of the French cooperation and their counterparts from Argentina, Brazil, Chile, Paraguay, Peru and Uruguay, aimed to promote and strengthen the collaboration and to create networks of research and development in the field of Information and Communication Science and Technology (ICT), through joint projects.

# SNPs-PHE Project

This project is oriented to the design of a knowledge-based model for the prediction and characterization of associations between Single Nucleotide Polymorphisms (SNPs) and phenotypic traits. Strategically, our proposal is based on the characterization of SNPs interactions by supervised machine learning techniques enriched and guided by expert knowledge, i.e., putative positions of mutations, candidate gene sets, inheritance issues, and epigenetic evidence. The broad characterization of SNPs interaction at gene and intergene levels is necessary to infer complex traits. The inclusion of expert knowledge is fundamental to reduce the computational complexity of such broad characterizations. Under this baseline, a prototype model for predicting eye color defective traits arising from SNPs in eye pigment genes of the fruit fly *Drosophila melanogaster* is considered. This project aims to contribute to a long-term objective of understanding complex traits, i.e., multigenic disorders. For this purpose, the development of machine learning methods for modeling expert knowledge and information sources are considered. In particular, the problem of predicting SNP-phenotypic trait associations involves noisy and multi-dimensionally correlated high dimensional datasets often lacking of the appropriate number of samples. The effective treatment of these information sources requires the integration of complex expert knowledge, a challenging issue to which the results of this project could bring an adequate solution.

# En menu Description:

# Image result for ommochrome eye pigment

Goals

The main goal of this project is the design of prediction model for SNPs and phenotypic trait associations able differentiation between two main types of SNPs interaction, redundant and complementary, along with a measure of their importance.  To this purpose, a prototype case of study based on the *Drosophila melanogaster* genomic model involving SNPs in eye pigment genes associated with the phenotypic trait “final eye pigments” will be used. Briefly, the brick-red eye colors is a result from the interaction between two pigment pathways, one produces red tones, and the other produces brown tones.  In addition, no-pigments entail white eyes. Mutations occurred in some of the involved genes, will produce changes in encoded enzymes that control the biosynthetic pathways and so, final color pigment is modified. This problem is complex enough to visualize the complexity of SNPs interactions and sufficiently studied to validate computational results.

Expected results

The prototype design of a prediction model for SNPs-phenotypic trait associations using fly model *Drosophila* melanogaster. More precisely, our goal is to strongly impact on our fields of research with new results that cover the following aspects:

* The identification of expert knowledge sources and definition of formal procedures for its acquisition. Both syntax and semantic aspects will be considered towards the posterior integration of such knowledge into prototype predictions models.
* Scalability approach for fuzzy measures identification of interacting information sources to characterize potential sets of interacting SNPs at gene and intergene levels.
* Integration of expert knowledge and machine learning methods under the common framework of set measures of fuzzy type.
* Validation of the proposal for predicting SNPs associated with phenotypic trait of eye pigments in *Drosophila melanogaster*.

Motivations and Methodology

The rapid progress in next generation sequencing (NGS) technology leads to a huge gap between data acquisition and data analysis. One of the main applications of NGS technologies is the identification of causative -no neutral- SNPs associated with phenotypic traits, with special interest in those leading to medical genetic disorders. A first way to distinguish deleterious from neutral SNPs is by performing case-control genome-wide association studies (GWAS) across populations. However, these studies do not solve the problem of deciding which SNPs are relevant to phenotypic traits, the holy grail of medical human genetics. Alternatively, machine learning-based models for predicting associations between SNPs and phenotypic traits have been considered. However, the complexity of dealing with thousands of thousands of SNPs together with the hardness of modeling the underlying genomics and systems biology, part of which is yet unknown,  makes the problem computationally hard. As a result, although simplified Support Vector Machine (SVM) prediction models considering raw SNPs processing can turn GWAS feasible, their prediction accuracy may not increase beyond chance. Only in few cases SNPs-phenotypic trait associations can be accurately predicted, e.g., in monogenic inheritance diseases with a 100% of penetrance. Actually, in most cases, SNPs-phenotypic trait associations involve multigenic disorders together with epigenetic and environmental factors affecting the penetrance degree. This complex knowledge is distributed across multiple sources, ranging from genetic experts to databases containing SNPs and -omics data illuminating enrichment studies. These considerations motivate our first working hypothesis:

*A proper characterization of SNPs coming from genomic projects together with the integration of expert knowledge by means of machine learning techniques may enhance current prediction models of  SNPs-phenotypic trait associations.*

Regarding expert knowledge, the impact of candidate small sets of SNPs associated with a target phenotypic trait is a main concern. A typical roadmap starts with the analysis of sets of SNPs at the gene and intergene level in well-defined genomic regions. For instance, to understand a genetic disorder produced by multigenic inheritance, candidate gene sets and associated SNPs are first identified. Gene sets are then characterized in terms of their linkage disequilibrium score (LD) indicating that involved genes are non-randomly associated. In addition, inheritance issues explained by mechanisms of epistasis indicating locus-locus interactions of multigenic disorders are further analyzed. Finally, epigenetic evidence indicating molecular modifications that are not set out in the DNA sequence level are also evaluated. The drawback of expert manual approach is the scalability: exhaustive analysis on N SNPs involves 2N set characterizations.

To overcome this problem, an initial dimensionality reduction of the SNPs search space is performed by using a feature selection technique based on genetic algorithms. In contrast to typical filtering techniques, this feature selection approach can take into account SNPs interactions when grouped together. Then we propose a two-stage computational approach based on, *a*) an expert-based dimensionality reduction of predefined candidate SNPs by the computational modeling of preferences about relevant genomic regions and maximum cardinality of relevant sets of interacting SNPs; preferences modeling mediated by a fuzzy measure framework, and *b*) a machine learning-based characterization of candidate sets of interacting SNPs in terms of their importance in the phenotypic trait. The importance of SNP subsets will be tackled by means of a scalable variant of fuzzy measures allowing the computation of an interaction index enablingthe characterization of SNPs redundancy and complementariness. From a biological point of view, a redundant set of SNPs could be interpreted as “any SNP within this set is responsible for a phenotypic trait” while a complementary set of SNPs could be interpreted as “as all the SNPs within this set are responsible for a phenotypic trait”. This approach may help to understand and quantify different types of associations between SNPs and phenotypic traits. These considerations motivate our second working hypothesis:

*A supervised fuzzy-measure characterization of sets of SNPs may contribute to better understand SNP architecture features associated with certain phenotypic traits.*

# En menu Members:

Argentina

**Research in Information Sciences, Bioinformatics & Agroinformatics, Universidad Nacional de Rosario – CIFASIS - CONICET**

Dra. Pilar Bulacio (International Coordinator) – bulacio http://www.fceia.unr.edu.ar/%7Ejcgomez/BAVI/html/arroba.pngcifasis-conicet.gov.ar

Dra. Flavia Krsticevic – krsticevic http://www.fceia.unr.edu.ar/%7Ejcgomez/BAVI/html/arroba.png cifasis-conicet.gov.ar

Dr. Javier Murillo – murillo http://www.fceia.unr.edu.ar/%7Ejcgomez/BAVI/html/arroba.png cifasis-conicet.gov.ar

Dra. Elizabeth Tapia – tapia http://www.fceia.unr.edu.ar/%7Ejcgomez/BAVI/html/arroba.png cifasis-conicet.gov.ar

France

**Université de Technologie de Compiegne, Centre de recherche de Royallieu**

Dr. Sebastien Destercke - sebastien.desterckehttp://www.fceia.unr.edu.ar/%7Ejcgomez/BAVI/html/arroba.pnghds.utc.fr

**Université Paris-Dauphine, Laboratoire d'Analyse et de Modélisation des Systèmes pour l'Aide à la Décision, LAMSADE**

Dr. Cailloux Olivier - olivier.caillouxhttp://www.fceia.unr.edu.ar/%7Ejcgomez/BAVI/html/arroba.pngdauphine.fr

**Institut national de recherche en sciences et technologies pour l'environnement et l'agriculture (IRSTEA).**

Dr. Serge Guillaume – serge.guillaumehttp://www.fceia.unr.edu.ar/%7Ejcgomez/BAVI/html/arroba.pngirstea.fr

Uruguay

**Departamento de Informática y Ciencias de la Computación, Universidad Católica del Uruguay**

Dr. Gustavo Vazquez - gustavo.vazquezhttp://www.fceia.unr.edu.ar/%7Ejcgomez/BAVI/html/arroba.pngucu.edu.uy

Lic. Tamara Fernandez – tamara.fernandezhttp://www.fceia.unr.edu.ar/%7Ejcgomez/BAVI/html/arroba.pngucu.edu.uy

|  |  |
| --- | --- |
| [http://www.cifasis-conicet.gov.ar/granitto/style/logo_cifasis.jpeg](http://www.cifasis-conicet.gov.ar/)C:\Users\pilar\AppData\Local\Microsoft\Windows Live Mail\WLMDSS.tmp\WLMEB4.tmp\logo UCU.png | [https://date.ac.lri.fr/img/cnrs.jpg](http://www.cnrs.fr/) |

[](http://www.irstea.fr/)

# UTC Sorbonne Universites

# [Université Paris Dauphine, retour à la page d'accueil](http://www.dauphine.fr/fr/index.html)

# En menu Publications:

# Publications

Flavio Spetale, Elizabeth Tapia, Javier Murillo, Flavia Krsticevic, Sergio Ponce, and Pilar Bulacio. *Proper integration of feature subsets boost GO subcellular localization predictions*. XXI Congreso Argentino de Bioingeniería y X Jornadas de Ingeniería Clínica, SABI 2017. Octubre 2017. In press.

Flavia Krsticevic, Flavio Spetale, Elizabeth Tapia, Javier Murillo, and Pilar Bulacio. *Predicting Cellular Component of Ommochrome Pathway eye genes in D. melanogaster based on Machine Learning*. X Simpósio de Ecologia, Genética e Evolução de Drosophila. Novembro 2017.

# En menu Activities:

# Workshop 1: Rosario, Argentina

# Meeting summary

# First day, Monday 17 April 2017

# Presentations:

# Single Nucleotide Polymorphisms Phenotype (Pilar Bulacio; Elizabeth Tapia)

# Phenotype eye color (Flavia Krsticevic)

# Second day, Tuesday 18 April 2017

# Presentations:

# Multicriteria Decision Aid, MCDA (Olivier Cailloux, Sébastien Destercke)

# Genetic algorithms for feature selection (Gustavo Vazquez)

# FlyVar-FlyBase discussion (Flavia Krsticevic, Tamara Fernandez)

# Third day, Wednesday 19 April 2017

# Presentations:

# Feature set interactions (Javier Murillo, Serge Guillaume)

# MCDA: Portfolio selection and case study (Olivier Cailloux)

# Dataset design: Discussion about data sources and processing

# Fourth day, Thursday 20 April 2017

# Presentations:

# Dataset design: Features and samples identification

# Fifth day, Friday 21 April 2017

# Presentations:

# Objective specification: pipeline definition

# Next meeting planning: 22-29 September, 2017 (France)



# Workshop 2: Compiègne, France

# Meeting summary: On going