GENS2: a Tool to Simulate Gene-gene and Gene-environment Interactions and Epistatsis

Scala G (1, 2), Pinelli M (1, 3), Amato R (1, 2), Miele G (1, 2, 4), Cocozza S (1, 3)
(1) Gruppo Interdipartimentale di Bioinformatica e Biologia Computazionale
Università di Napoli Federico II - Università di Salerno
(2) Dipartimento di Scienze Fisiche, Università di Napoli Federico II, Napoli
(3) Dipartimento di Biologia e Patologia Cellulare e Molecolare L. Califano
Università di Napoli Federico II, Napoli
(4) INFN Sezione di Napoli, Complesso Universitario di Monte S.Angelo, Napoli

Motivation
Complex diseases (CD) are multifactorial traits caused by both genetic and environmental factors. They represent the most part of human diseases and include those with largest prevalence and mortality (cancer, heart disease, obesity, etc). The concept of Gene-Gene (GxG) and Gene-Environment interaction (GxE) is theoretically central in CD, and it is widely accepted that these phenomena should be considered to avoid a serious underestimation of the disease risk and inconsistencies of replication among different studies. Furthermore, taking into account the GxG and GxE could focus medical intervention by identifying sub-groups of individuals who are more susceptible to specific environmental exposures. However, despite of a large amount of information that have been collected about both genetic and environmental risk factors, there are relatively few examples of studies on their interactions in epidemiological literature. One reason can be the incomplete knowledge of the power of statistical methods designed to search for risk factors and their interactions in these data sets. A possible strategy to overcome this limitation can be to challenge the different statistical methods against data sets where the underlying phenomenon is completely known and fully controllable like, for example, simulated ones. In this direction, we developed a mathematical model (Multi-Logistic Model, MLM) to describe a wide variety of gene-environment interactions.

Methods
The multi-logistic model is based on a series of logistic relationships, one for each combination of genotypes, in which the disease risk is a function of the environmental exposure. The MLM can model any type of gene-gene and gene-environment interactions, involving any number of genetic and environmental factors and also allowing non-linear interactions as epistasis. However, it is still difficult to specify the parameters needed by
the MLM in order to simulate biologically meaningful and understandable conditions. We faced this problem by designing a subsystem, called Knowledge Aided Parameterization System 2 (KAPS2), which translates epidemiological and biological information in a set of parameters to input into the MLM. By KAPS2 is currently possible to model interactions among up to 2 genes and 1 environmental factor, also considering non-linear gene-gene interactions like epistasis. In our model the two genetic factors are basically considered as independently influencing the disease risk and the epistasis as a departure from that condition of independence. Moreover, to describe the interaction between the genetics and the environment, we propose two main interaction models. The Additive Model (AM), where genetic and the environment influence the risk directly, independently and additively; and the Gene Environment interaction Model (GEM) where the genetics does not directly affect the disease risk, but modulates the response to the environmental exposure. These models are translated into parameters for the MLM by a non-linear system of equations that reflects the relationship between standard epidemiological measures and MLM parameters. Regarding the epistasis, the departure from the independent interaction between the two genes, is modeled as a set of linear constraints that define admissible disease risk variations for each combination of genotypes and is completely determined using a non linear optimization algorithm.

Results
We implemented the MLM and the KAPS2 system in a tool named GENS2 that can produce case-control samples of complex disease. Furthermore, a Monte Carlo process allows a random variability in the results, which can be particularly relevant to test statistical methods. GENS2 was designed as a module for the SimuPOP simulation framework, which can simulate whole-genome data sets with realistic linkage disequilibrium patterns.

Contact email
giovanni.scala2@studenti.unina.it; michele.pinelli@unina.it