Challenging an ensemble approach (GENTES) with the Gene-Environment iNteraction Simulator (GENS)

D'Andrea D[1], Amato R[1,2,3], Pinelli M[1,4], Tagliaferri R[1,5], Cocozza S[1,4], Miele G[1,2,3]

Motivation

Complex diseases are multifactorial traits caused by both genetic and environmental factors. They represent the major part of human diseases and include those with largest prevalence and mortality (cancer, heart disease, obesity, etc.). Despite a large amount of information that has been collected about both genetic and environmental risk factors, there are few examples of studies on their interactions in epidemiological literature. One reason can be the incomplete knowledge of the statistical power of Feature Selection Method (FSM) usually used to identify the risk factors and their interactions in data sets. As it is well known, each FSM have different performances and weaknesses and better performs in particular conditions. It's clear that an improvement in this direction would lead to a better understanding and description of gene-environment interactions. To this aim, a possible strategy is to challenge the different statistical methods against data sets where the underlying phenomenon is completely known and fully controllable, for example simulated ones; to determine rules to improve performances of each method; combining FSMs in an ensemble approach to add the positive characteristics of each method and to dilute at the same time the weakness points.

Methods

We present a mathematical approach that models gene-environment interactions. By this method it is possible to generate simulated populations having gene-environment interactions of any form, involving any number of genetic and environmental factors and also allowing non-linear interactions as epistasis. In particular, we implemented a simple version of this model in a Gene-Environment iNteraction Simulator (GENS), a tool designed to simulate case-control data sets where a one gene-one environment interaction influences the disease risk. The main aim has been to allow the input of population characteristics by using standard epidemiological measures and to implement constraints to make the simulator behavior biologically meaningful. Then we developed new software, Gene-Environment iNteraction Exploration System (GENTES), that implements an

[1] Gruppo Interdipartimentale di Bioinformatica e Biologia Computazionale, Università di Napoli "Federico II" - Università di Salerno, Italy. [2] Dipartimento di Scienze Fisiche, Università di Napoli "Federico II", Napoli, Italy. [3] INFN Sezione di Napoli, Napoli, Italy. [4] Dipartimento di Biologia e Patologia Cellulare e Molecolare "L. Califano", Napoli, Italy. [5] Dipartimento di Matematica e Informatica, Università di Salerno, Fisciano (SA), Italy

ensemble of FSMs aimed to identify relevant genetic and non-genetic features involved in a given complex disease. The ensemble can be composed by any type of FSM, as well we present the implementation of four of them, namely the Binary Logistic Regression, the Linear Discriminant Analysis, the Multifactor Dimensionality Reduction and an univariate $2 or t-test. We optimized the performances of the ensemble in identifying gene-environment interactions by the challenges on simulated datasets.

Results

By the multi-logistic model implemented in GENS it is possible to simulate case-control samples of complex disease where gene-environment interactions influence the disease risk. The user has full control of the main characteristics of the simulated population and a Monte Carlo process allows random variability. A knowledge-based approach reduces the complexity of the mathematical model by using reasonable biological constraints and makes the simulation more understandable in biological terms. Simulated data sets were used to evaluate the statistical power of four widely used FSM. Moreover the same simulated data sets were used to evaluate the performances of the ensemble in comparison with those of single FSMs. The ensemble showed performances generally better than or comparable to those of each one of its components.

Contact e-mail

daniel.dandrea@gmail.com