

Analisi della relazione genotipo fenotipo con algoritmi di Machine Learning

Progetto di Ricerca

Insieme allo stile di vita, il contributo genetico ha un ruolo importante nel determinare la predisposizione a patologie complesse. In anni recenti, la diffusione delle tecnologie di sequenziamento del DNA ha portato ad una crescente disponibilità di informazione genetica. Oggi, uno dei principali problemi della genetica umana è come collegare i dati genetici alle caratteristiche fenotipiche. Gli algoritmi di Machine Learning hanno la capacità di apprendere autonomamente relazioni tra dati, e conseguentemente sono ritenuti un approccio promettente per estrarre informazioni clinicamente rilevanti a partire dalle informazioni genetiche. L'obiettivo del progetto di ricerca è sviluppare metodi interpretabili per la predizione di caratteristiche fenotipiche a partire da dati di varianti genomiche.

Piano delle Attività

L'utilizzo di metodi di Machine Learning per l'analisi della relazione genotipo-fenotipo è attualmente ostacolato da due problemi principali: (1) la numerosità dei campioni disponibili rispetto alla dimensione dello spazio delle caratteristiche genetiche; (2) l'interpretabilità dei risultati ottenuti in termini medico/biologici. Per provare a mitigare entrambi i problemi sarà testato l'utilizzo di reti neurali che integrano nella loro architettura informazioni biologiche a priori. In questo modo sarà possibile contemporaneamente ridurre il numero di parametri da determinare nella fase di training, e ottenere delle variabili latenti direttamente collegate a concetti biologici noti. Questo tipo di approccio ha la potenzialità di rendere il dato genetico più facilmente utilizzabile nella pratica clinica.

Analyses of genotype-phenotype relation by Machine Learning methods

Research Project

In combination with lifestyle, genetics plays an important role in determining the predisposition to complex diseases. In recent years, thanks to DNA sequencing technologies, the availability of genetic data has been rapidly increasing. Today, one of the major problems in human genetics is how to link genotype information to phenotypic characteristics. Machine Learning algorithms have the ability to autonomously learn relationships between data, and are consequently considered a promising approach for extracting clinically relevant information from genetic data. The objective of this research project is to develop interpretable methods for the prediction of phenotypic characteristics starting from genomic variants.

Plan of Activities

The use of Machine Learning methods for the analysis of the genotype-phenotype relation is currently hindered by two main problems: (1) the size of the datasets compared to the high dimensionality of the genetic input features; (2) the interpretability of the results obtained in medical/biological terms. To try to mitigate both problems, the use of neural networks that incorporate prior biological information into their architecture will be tested. In this way it will be possible to simultaneously reduce the number of parameters to be estimated in the training phase, and to define latent features that are directly linked to known biological items. This approach has the potential to make the information extracted from genetic data more easily adoptable in the clinical practice.