Efficient algorithm analyzes large genotypic and phenotypic data sets in clinical trials

Published date: Feb. 1, 2012

Technology description

Summary

MARKETS ADDRESSED:
The algorithm has a wide range of practical uses in many fields where the goal is to identify associations between high-dimensional predictor combinations (genotype) and responses (phenotype). One example is the investigation of the association between genotype and genetic sequence combinations (e.g. from microarrays, SNPs, etc.) and one or more phenotypes of interest. The inventors current application with the algorithm is the for the selection of semi parametric regression models for the association between multi-way combinations of HIV-1 genotype codon/amino acid pairs and clinical responses (phenotypes). Examples of phenotypes include the risk of future treatment failure, clinical events and drug toxicities. Also of interest is the investigation of genotype time evolution among patients experiencing virologic failure.
Recent technologies such as genotype sequencing and gene expression arrays result in datasets that typically have many more variables than data points. As a consequence, these variables have a complex, high-dimensional dependency structure that is most likely unknown. As such, practical methods for investigating the association between multi-way combinations of multiple predictor variables (genotype) and one or more phenotypes may be useful in many areas of scientific research. For example, estimating the combinations of HIV-1 genotype codon/amino acid pairs that are associated with future HIV-1 RNA response may be potentially useful for patient-specific treatment management. When attacking such problems the number of potential genotype combinations is so enormous that current methods do not allow such an analysis. A novel and efficient software algorithm that makes it feasible to analyze such problems in practice. In one example, the method was shown to perform well in a 24 week AIDS clinical trial determining the association between HIV-1 RNA level and genotype sequence weeks after study entry. In the study, an HIV-1 genotype sequence from the protease and reverse transcriptase regions was obtained for all subjects at baseline and at several follow-up visits for those subjects whose current HIV-1 RNA level was large enough to permit amplification. The variable at each position (codon) in the genotype sequence is an amino acid, taking one of 20 possible unordered values. A 24 week HIV-1 RNA value was determined to reflect important information about a future clinical response, particularly information on specific HIV-1 drugs that a

subject's viral population is thought to manifest resistance in the HIV-1 genotype sequence.
This method is clever and not obvious to the ordinary biostatistician. The ordinary analyst may attempt to use current tools for high-dimensional regression, such as the elements of statistical learning--for example, ridge regression, Lasso, and Least angle regression directly on the full dataset of phenotype(s) and all possible genotype combinations. The current approach also translates a computer-memory intensive analysis into a CPU-intensive analysis, making the problem more tractable in practice.
The method also does not put any restrictions on the data-generating distribution and thus is useful for a wide range of applications. Sparse data is also naturally accommodated. The methodology provides control over the number or proportion of false positive results. The number of models in this set is easily controlled. Since genotype assays have dropped to a cost that permits their routine use, an understanding of the association between HIV-1 genotype and RNA could be used by clinicians to help in the selection of patient-specific drug regimens.

Institution

Harvard University