Classifier Ensemble Based Analysis of a Genome-Wide SNP Dataset Concerning Late-Onset Alzheimer Disease

Lúcio Coelho, Biomind LLC, USA

Ben Goertzel, Biomind LLC, USA, and Xiamen University, China

Cassio Pennachin, Biomind LLC, USA

Chris Heward, Kronos Science Laboratory, USA

ABSTRACT

In this paper, the OpenBiomind toolkit is used to apply GA, GP, and local search methods to analyze a large SNP dataset concerning Late-Onset Alzheimer's Disease (LOAD). Classification models identifying LOAD with statistically significant accuracy are identified as well as using ensemble-based important features analysis in order to identify brain genes related to LOAD, most notably the solute carrier gene SLC6A15. Ensemble analysis is used to identify potentially significant interactions between genes in the context of LOAD.

Keywords: *Alzheimer's, Genetic Algorithms, Genetic Programming, Late-Onset Alzheimer's Disease (LOAD), Single-Nucleotide Polymorphisms*

INTRODUCTION

We report results obtained by using the OpenBiomind machine learning based bioinformatics data analysis toolkit (see http://code.google.com/p/openbiomind/) for the analysis of a dataset of Single-Nucleotide Polymorphisms (SNPs) concerning late-onset Alzheimer's Disease (LOAD) (Reiman et al., 2007). Putting it simply, SNPs are genomic variations of a single DNA base observed in the population under study. Although they not necessarily have a direct relation with a given phenotype (say, a medical condition with some potential genetic background), they often serve as markers commonly associated with those phenotypes. Hence the interest in studying SNPs for the development of diagnosis tools or for collecting clues that might point to genes interesting for shedding light into the problem at hand.

DOI: 10.4018/jssci.2010100105

We used the OpenBiomind functions which provide Genetic Algorithms (GA), Genetic Programming (GP) and local search methods for supervised categorization; and also functions which enable several forms of ensemble-based analysis, including important features analysis, Model Utilization-Based Clustering (MUTIC) and Model Based Role Analysis (MOBRA). In this way we:

- Obtained classification rules that can distinguish LOAD from Control via SNP combinations.
- Identified a number of potentially LOAD related SNPs and genes.

Classification results were tested with permutation analysis and the accuracy results are described here. While the results are highly statistically significant, the accuracy is not sufficient to serve as the basis of a practical diagnostic test. Further, it is our tentative, subjective opinion based on this work that no analytic method is going to be able to figure out a practical diagnostic test based on the SNPs in this study.

However, regarding the identification of genes potentially related to Alzheimer's Disease (AD), our results are more promising. The metatask classification method used by OpenBiomind provided a set of most important (in the context of LOAD) SNPs, and their related genes. The relations of these genes to LOAD were assessed by clustering and with biological analyses. Visualization for clustering results and SNP inter-relations are shown in order to ease the interpretation. Results show a few brain genes that we believe have a high chance of being related to LOAD, especially the solute carrier gene, SLC6A15.

The results presented here constitute another piece of evidence in favor of the power of the unorthodox OpenBiomind methods to yield statistically meaningful analytical results and novel, suggestive biological hypotheses.

DATA

Original Data

The dataset used in this study comprises the genome-wide SNP mapping of 1411 samples. 859 of them are case for late-onset Alzheimer Disease (LOAD), while the remaining 552 are control. Each sample is characterized by 312,316 single-nucleotide polymorphisms (SNPs). A more thorough description is given in Reiman et al. (2007).

The samples in the dataset are divided into three different cohorts representing different sources: a "Neuropathological Discovery Cohort" of 736 brain donors, a "Neuropathological Replication Cohort" of 311 brain donors and an additional "Clinical Replication Cohort" of 364 living subjects.

Sample Partition and SNP Selection

Machine learning experiments reported here were executed over the following sample partitions:

- All samples.
- Neuropathological discovery cohort.
- Neuropathological replication cohort.
- Clinical replication cohort.

Each of the sample partitions above was by its turn divided in three pairs of train-test sets, in a 3-fold cross-validation. All classification results reported here were obtained by applying classification methods to those folds divisions.

Finally, for each of those sample partitions, the following kinds of SNP selection were performed:

- **Homozygosis-based selection:** selection of the top 100 SNPs with highest positive difference of homozygosis frequency in case and control samples. In other words, selected SNPs are those with high homo-