Prediction of genetic risk of complex diseases by supervised learning

Vincent Botta VINCENT.BOTTA@ULG.AC.BE
Pierre Geurts P.GEURTS@ULG.AC.BE
Louis Wehenkel L.WEHENKEL@ULG.AC.BE
Department of Electrical Engineering and Computer Science
GIGA-Research, University of Liège, B4000 Belgium

Sarah Hansoul S.HANSOUL@ULG.AC.BE
Animal Genomics
GIGA-Research, University of Liège, B4000 Belgium

1. Whole genome association studies

The majority of important medical disorders (f.i. susceptibility to cancer, cardiovascular diseases, diabetes, Crohn's disease) are said to be complex. This means that these diseases are influenced by multiple, often interacting environmental and genetic risk factors. The fact that individuals differ in terms of exposure to environmental as well as genetic factors explains the observed inter-individual variation in disease outcome (i.e. phenotype). The proportion of the phenotypic variance that is due to genetic factors (heritability) typically ranges from less than 10 to over 60 % for the traits of interest. The identification of genes influencing susceptibility to complex traits reveals novel targets for drug development, and allows for the implementation of strategies towards personalized medicine.

Recent advances in marker genotyping technology allow for the genotyping of hundreds of thousands of Single Nucleotide Polymorphisms (SNPs) per individual at less than 0.1 eurocents per genotype, the identification of genomic regions (i.e. loci) that influence susceptibility to a given disease can now be obtained by means of so-called "whole genome association studies" (WGAS).

2. Supervised learning for WGAS

The basic idea behind a GWAS is to genotype a collection of affected (cases) and unaffected (controls) individuals for a very large number of SNPs spread over the entire genome. Genomic regions showing statistical differences among cases and controls are then detected using this dense collection of SNPs. From a machine learning point of view, analysis of this dataset is a binary classification problem, with a very large number of raw symbolic variables, each one corresponding to a different SNP and having only three possible values (homozygous wild, heterozygous and homozygous mutant). On top of this very high p/n ratio, these problems are also generally highly noisy, and the raw input variables are strongly correlated (which is explained by the so-called linkage desiquilibrium).

In this research we study two different representations of the input data for the application of supervised learning, namely the raw genotype data on the one hand, and on the other hand the groups of strongly correlated SNPs (i.e. the haplotype blocks) , representing the observed combinations of about 10 to 100 phased genotypes between the recombination hotspots of the different chromosomes. We report an empirical study based on several simulated datasets where one or two independent or interacting causal mutations on a single chromosome are studied. We provide comparative results of different ensembles of randomized decisions trees adapted to handle the particular nature of these two types of input variables. These methods are assessed in terms of their predictive power as well as their ability to help identifying the genomic regions containing causal mutations.

3. Methods

3.1. Dataset generation

We used the program *gs* (Li & Chen, 2008) to generate samples based on *HapMap* data (Consortium, 2003) so as to keep the linkage desiquilibrium patterns similar to those in actual human populations and focus on chromosome 5. The raw input variables were obtained by taking SNPs spaced by 10 kilobases from the *HapMap* pool to reproduce classical GWAS conditions, and the causal disease loci were removed from the input variables.

Using genotype penetrance tables, 5 different disease models were tested: two for the one locus experiments,

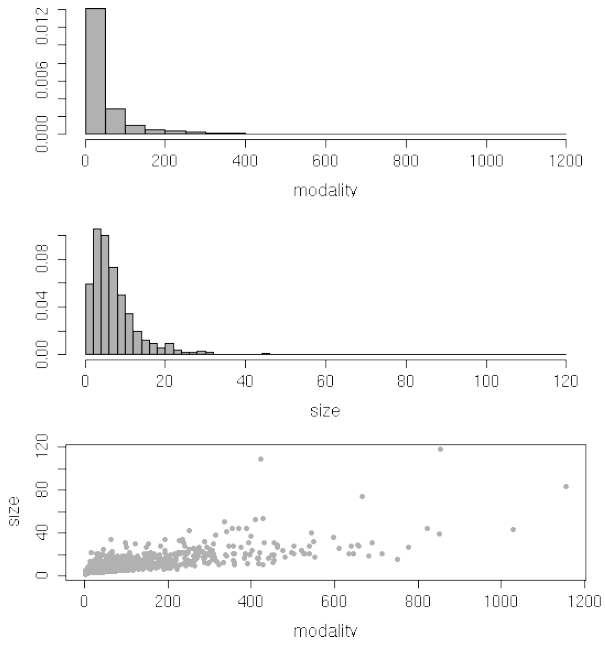

Figure 1. Haplotype block statistics. Top: block modalities; Middle: block length; Bottom: scatter plot

and the three most common disease models with interactions (Li & Reich, 2000) for the two locus case. We considered different noise and penetrance values.

In the first (raw) data representation, the different databases are composed of 14604 symbolic variables with 3 possible values. The second representation is a variant of the first where we group correlated variables into blocks (haplotype blocks chosen according to *HapMap* hotspots). This dramatically reduces the number of variables but it also increases their modalities (up to few hundred possible combinations when the sample comes from a broad population). In total, this yielded 1957 haplotype blocks. Figure 1 shows the histograms of block lengths and number of modalities.

3.2. Supervised learning

We evaluated Random Forests and Extra-Trees (see Geurts et al., 2006 for a precise description of these algorithms and related notions). These methods were customized in an ad hoc way to handle the datasets for the haplotype block variant. Various values of their two main meta-parameters (number of tested attributes and number of trees) were screened while the trees were completely developped.

Learning was repeated 10 times on balanced learning sets (containing between 100 and 1000 controls and as many cases). All models were evaluated on the same independent and balanced test set of size 5000.

The predictive power was assessed using the mean area under the ROC curves and compared to best possible theoretical AUCs which were deduced from the selected disease model.

We ranked SNPs and haplotype blocks using variable importances based on information theory (see Wehenkel, 1998), and provide the mean rank of the SNPs adjacent to the causal mutations, or of the block(s) containing these mutation(s).

4. Preliminary results

Preliminary results show good perspectives. In particular, the different methods obtain rather good AUCs as compared with the theoretical upper bound derived from the disease models. The different methods are also able to predict and to localize the disease loci, rather well. We also observed that most often the direct application of supervised learning to the raw genotype data provides slightly superior results both in terms of risk prediction and loci identification than the application of these methods to haplotype blocks. This essentially suggests that further work should focus on a better determination of the haplotype block structure from the datasets themselves (rather than by extrapolating these structures from other cohorts, as it was the case in these first investigations).

Acknowledgments

This paper presents research results of the Belgian Network BIOMAGNET (Bioinformatics and Modeling: from Genomes to Networks), funded by the Interuniversity Attraction Poles Programme, initiated by the Belgian State, Science Policy Office. The scientific responsibility rests with its authors. Vincent Botta is recipient of a F.R.I.A. fellowship. Sarah Hansoul is a postdoctoral research fellow of the F.R.S.-FNRS and Pierre Geurts is a Research Associates of the F.R.S.-FNRS.

References

Consortium, T. I. H. (2003). The international hapmap project. *Nature*, *426*, 789–796.

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees.

Li, J., & Chen, Y. (2008). Generating samples for association studies based on hapmap data. *BMC Bioinformatics*, *9*, 44.

Li, W., & Reich, J. (2000). A complete enumeration and classification of two-locus disease models. *Hum Hered*, *50*, 334–349.

Wehenkel, L. (1998). Automatic learning techniques in power systems.