Laboratory Note

Genetic Epistasis
VIII - Assessing Algorithm MBMDR

LN-8-2014

Ricardo Pinho and Rui Camacho
FEUP
Rua Dr Roberto Frias, s/n,
4200-465 PORTO
Portugal
Fax: (+351) 22 508 1440
e-mail: ei09045@fe.up.pt
www : http://www.fe.up.pt/∼ei09045
rcamacho@fe.up.pt
www : http://www.fe.up.pt/∼rcamacho

May 2014

Abstract

Model-Based Multifactor DImensionality Reduction (MBMDR) is an algorithm that implements on the previous MDR methodology, which consists on dividing SNPs into two clusters based on their risk to the determination of the disease. Instead of using a predetermined threshold from the frequency of SNPs in the data, MBMDR uses a testing approach followed by a significance assessment. The results show a high Power only for large sized data sets and very low Type 1 Error Rate for all configurations. The running time of the algorithm makes the algorithm not viable for larger data sets.

1 Introduction

Multifactor DImensionality Reduction (MDR) [CLEP07] is one of the most referenced algorithms for epistasis detection. MDR filters SNPs, based on the frequency in case control data, to divide SNPs into high risk or low risk based on a predetermined threshold. Using cross validation and permutations to determine the high/low risk groups, the algorithm returns the high risk loci that have a stronger connection in the disease outcome. However, it samples many SNPs together analysing at most one significant epistasis model, skiping other possible SNP groups that may not have such significant conection, but may also be related to the disease.

Model-Based Multifactor Dimensionality Reduction [MVV11] merges multi-locus genotypes that have significant high or low risk based on association testing, rather than a threshold value.

MB-MDR process can be divided into the following steps:

1. **Multi-locus cell prioritization** - Each two-locus genotype is assigned to either High risk, Low risk or No Evidence of risk categories.

2. **Association test on lower-dimensional construct** - The result of the first step creates a new variable with a value correlated to one of the categories. This new variable is then compared with the original label to find the weight of high and low risk genotype cells.

3. **Significance assessment** - This stage tries to correct the inflation of type I errors after the combination of cells into the weight of High risk and Low risk. This is done using the Wald statistic.

1.1 Input files

The input file consists of the Index and phenotype in the first two columns, and the genotype of each SNP in the following columns. The first row corresponds to the name of each column.

```
””,”Y”,”SNP1”, ”SNP2”, ”SNP3”, ”SNP4”, ”SNP5”
”0”, 0,   1,      2,      0,      0,      0
”1”, 0,   0,      2,      1,      2,      0
”2”, 1,   1,      0,      1,      0,      1
”3”, 1,   1,      1,      2,      1,      0
```

Table 1: An example of the input file containing genotype and phenotype information with 5 SNPs and 4 individuals.

1.2 Output files

The output consist of a list of SNP interactions selected with the following columns for each interaction:

1. **SNP1...SNPx** - Names of snps in interaction.
2. **NH** - Number of significant High risk genotypes in the interaction.
3. **betaH** - Regresion coeficient in step2 for High risk exposition.
4. **WH** - Wald statistic for High risk category.
5. **PH** - P-value of the Wald test for the High risk category.
6. **NL** - Number of significant Low risk genotypes in the interaction.
7. **betaL** - Regresion coeficient in step2 for Low risk exposition.
8. **WL** - Wald statistic for Low risk category.
9. **PL** - P-value of the Wald test for the Low risk category.
10. **MIN.P** - Minimun p-value (min(PH,PL)) for the interaction model.

1.3 Parameters

The MBMDR can contain the following arguments:

- *order* - dimension of interactions to be analyzed.
- *covar* - (Optional) a data frame containing the covariates for adjusting regression models.
- *exclude* - (Optional) Value/s of missing data.
- *risk.threshold* - Threshold used to define the risk category of a multilocus genotype. The default value is 0.1.
- *adjust* - (Optional) Types of regression adjustment. Can be "none", "covariates", "main effects" or "both". The default value is "none".
- *first.model* - Specifies the first interaction to be tested. Useful when stoped before finishing the complete analysis.
- *list.models* - (Optional) Exhaustive list of models to be analyzed. Only possible interactions in this list will be analyzed.

- *use.logistf* - Boolean value indicating wheter or not the logistf package should be used. The default value is TRUE.
- *printStep*1 - Boolean value that prints every model obtained if the value is TRUE. The default value is FALSE.

2 Experimental Settings

The datasets used in the experiments are characterized in Lab Note 1.
The limit number of interactions selected is 2, considering that the ground truth is a pairwise interaction, and all of the SNPs are tested with each other for pairwise interactions.

3 Results

The algorithm only outputs the statistical relevancy test of interactions between SNPs. Due to this, only epistatic disease model data sets will be used for this experiment. Because of time constraints, several computers were used to obtain results. This means that it is not possible to compare scalability results.
The Figure 1 reveals a large increase with population size for data sets with a minor allele frequency higher than 0.01. There is a big increase in data sets with 2000 individuals from a minor allele frequency of 0.05 to 0.1. The results from data sets with a smaller amount of population size has much lower Power, having 0 Power for almost all data sets with 500 individuals. There is also a clear increase with minor allele frequency.

According to Figure 2 the Type 1 Error Rate is very low across all allele frequencies and data set sizes, having a maximum of 6% and 2% for 0.05 minor allele frequency with 2000 and 1000 individuals respectively. For other allele frequencies, only 0.1 and 0.3 contain false positives for data sets with 2000 individuals.

Figure 3 and 6 show the same results as Figure 1, with a different prespective. Figure 4 also shows an increase in Power with the increase in odds ratio. Figure 5 shows a smaller increase in Power with the increase in prevalence.

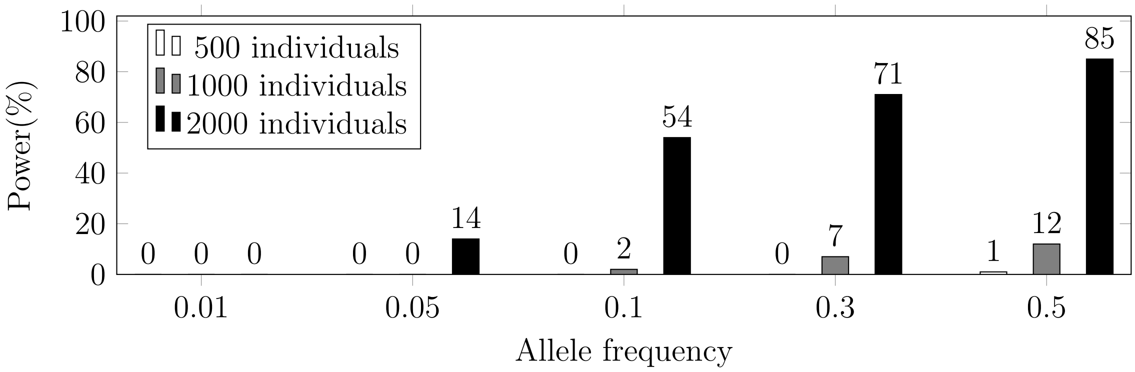

Figure 1: Power by allele frequency. For each frequency, three sizes of data sets were used to measure the Power, with an odds ratio of 2.0 and prevalence of 0.02. The Power is measured by the amount of data sets where the ground truth was amongst the most relevant results, out of all 100 data sets.

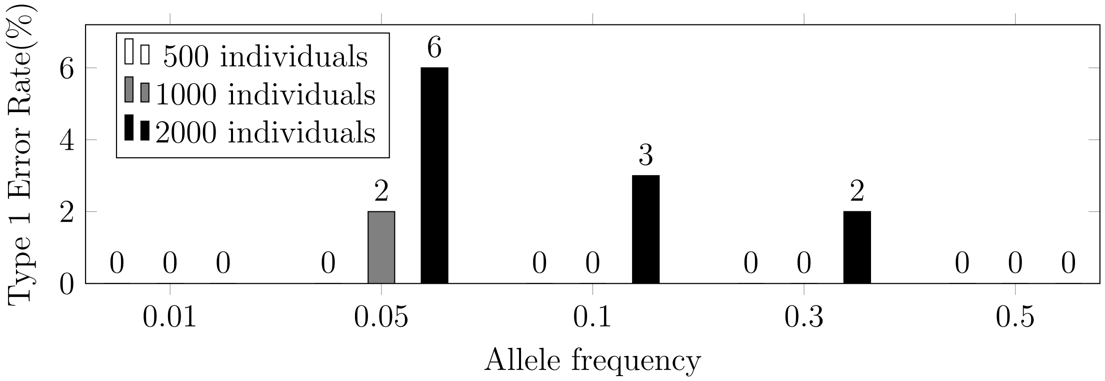

Figure 2: Type 1 Error Rate by allele frequency and population size. The Type 1 Error Rate is measured by the amount of data sets where the false positives were amongst the most relevant results, out of all 100 data sets.

4 Summary

MBMDR is an algorithm based on the popular MDR approach, with a clustering of SNPs by high and low risk of determining the disease phenotype. The results show very high Power for data sets with 2000 individuals, but very low Power for all other configurations. The Type 1 Error Rate is very low, reaching a maximum of only 6% for 0.05 allele frequency and 2000 individuals. Considering that there are no results concerning the scalability due to the expected running time of the algorithm shows that it is not viable to use this algorithm on big data sets that might contain thousands or millions of SNPs.

References

[CLEP07] Yujin Chung, Seung Yeoun Lee, Robert C Elston, and Taesung Park. Odds ratio based multifactor-dimensionality reduction method for detecting gene-gene interactions. *Bioinformatics (Oxford, England)*, 23:71–76, 2007.

[MVV11] Jestinah M Mahachie John, Francois Van Lishout, and Kristel Van Steen. Model-Based Multifactor Dimensionality Reduction to detect epistasis for quantitative traits in the presence of error-free and noisy data. *Eur J Hum Genet*, 19(6):696–703, June 2011.

A Bar Graphs

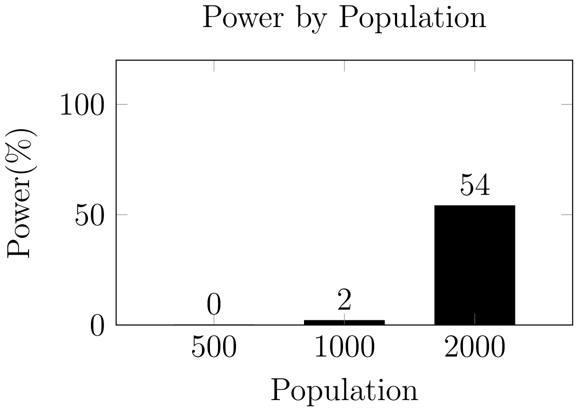

Figure 3: Distribution of the Power by population. The allele frequency is 0.1, the odds ratio is 2.0, and the prevalence is 0.02.

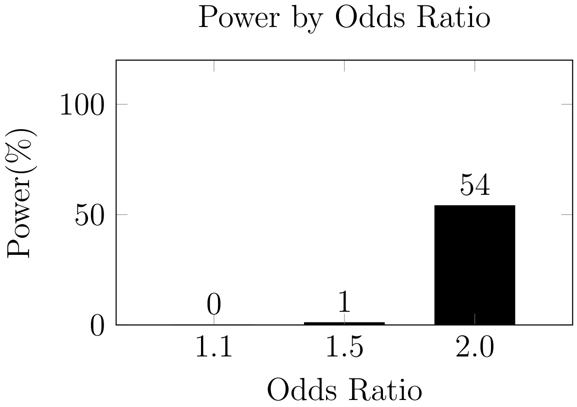

Figure 4: Distribution of the Power by odds ratios. The allele frequency is 0.1, the number of individuals is 2000, and the prevalence is 0.02.

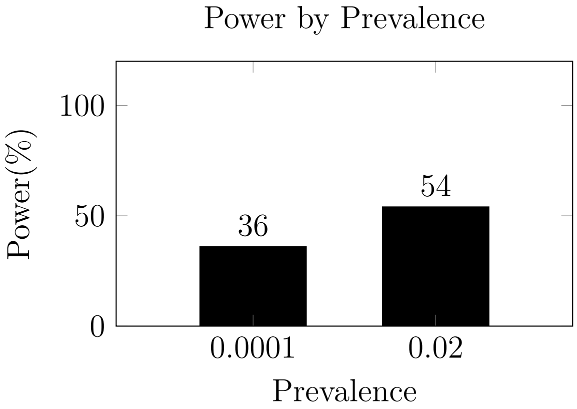

Figure 5: Distribution of the Power by prevalence. The allele frequency is 0.1, the number of individuals is 2000, and the odds ratio is 2.0.

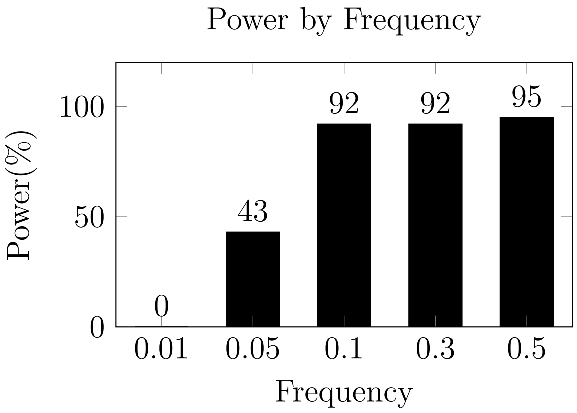

Figure 6: Distribution of the Power by allele frequency. The number of individuals is 2000, the odds ratio is 2.0, and the prevalence is 0.02.

B Table of Results

Table 2: A table containing the percentage of true positives and false positives in each configuration. The first column contains the description of the configuration. The second and third columns contain the number of datasets with true positives and false positives respectively, out of all 100 data sets per configuration.

Configuration*	TP (%)	FP (%)
0.5,500,I,2.0,0.02	1	0
0.5,500,I,2.0,0.0001	2	0
0.5,500,I,1.5,0.0001	0	0
0.5,500,I,1.1,0.02	0	0
0.5,500,I,1.1,0.0001	0	0
0.5,2000,I,2.0,0.02	85	0
0.5,2000,I,2.0,0.0001	91	2
0.5,2000,I,1.5,0.02	17	1
0.5,2000,I,1.5,0.0001	2	0
0.5,2000,I,1.1,0.02	0	0
0.5,2000,I,1.1,0.0001	0	0
0.5,1000,I,2.0,0.02	12	0
0.5,1000,I,2.0,0.0001	26	0
0.5,1000,I,1.5,0.02	0	0
0.5,1000,I,1.5,0.0001	0	10
0.3,500,I,2.0,0.02	0	0
0.3,500,I,2.0,0.0001	11	0
0.3,500,I,1.5,0.02	0	0
0.3,500,I,1.5,0.0001	0	0
0.3,500,I,1.1,0.02	0	0
0.3,500,I,1.1,0.0001	0	0
0.3,2000,I,2.0,0.02	71	2
0.3,2000,I,2.0,0.0001	100	8
0.3,2000,I,1.5,0.02	5	0
0.3,2000,I,1.5,0.0001	43	2
0.3,2000,I,1.1,0.02	0	0
0.3,2000,I,1.1,0.0001	0	0
0.3,1000,I,2.0,0.02	7	0
0.3,1000,I,2.0,0.0001	62	0
0.3,1000,I,1.5,0.02	0	0
0.3,1000,I,1.5,0.0001	5	0

0.3,1000,I,1.1,0.02	0	0
0.3,1000,I,1.1,0.0001	0	0
0.1,500,I,2.0,0.02	0	0
0.1,500,I,2.0,0.0001	0	0
0.1,500,I,1.5,0.02	0	0
0.1,500,I,1.5,0.0001	0	0
0.1,500,I,1.1,0.02	0	0
0.1,500,I,1.1,0.0001	0	0
0.1,2000,I,2.0,0.02	54	3
0.1,2000,I,2.0,0.0001	36	2
0.1,2000,I,1.5,0.02	1	0
0.1,2000,I,1.5,0.0001	0	0
0.1,2000,I,1.1,0.02	0	0
0.1,2000,I,1.1,0.0001	0	0
0.1,1000,I,2.0,0.02	2	0
0.1,1000,I,2.0,0.0001	1	0
0.1,1000,I,1.5,0.02	0	0
0.1,1000,I,1.5,0.0001	0	0
0.1,1000,I,1.1,0.02	0	0
0.1,1000,I,1.1,0.0001	0	0
0.05,500,I,2.0,0.02	0	0
0.05,500,I,2.0,0.0001	0	0
0.05,500,I,1.5,0.02	0	0
0.05,500,I,1.5,0.0001	0	0
0.05,500,I,1.1,0.02	0	0
0.05,500,I,1.1,0.0001	0	0
0.05,2000,I,2.0,0.02	14	6
0.05,2000,I,2.0,0.0001	3	1
0.05,2000,I,1.5,0.02	7	3
0.05,2000,I,1.5,0.0001	17	7
0.05,2000,I,1.1,0.02	0	0
0.05,2000,I,1.1,0.0001	0	0
0.05,1000,I,2.0,0.02	0	2
0.05,1000,I,2.0,0.0001	0	0
0.05,1000,I,1.5,0.02	0	0
0.05,1000,I,1.5,0.0001	0	0
0.05,1000,I,1.1,0.02	0	0
0.05,1000,I,1.1,0.0001	0	1

0.01,500,I,2.0,0.02	0	0
0.01,500,I,2.0,0.0001	0	0
0.01,500,I,1.5,0.02	0	0
0.01,500,I,1.5,0.0001	0	1
0.01,500,I,1.1,0.02	0	0
0.01,500,I,1.1,0.0001	0	0
0.01,1000,I,1.5,0.0001	0	0
0.01,1000,I,1.1,0.0001	0	0

*MAF,POP,MOD,OR,PREV where MAF represents the minor allele frequency, POP is the number of individuals, MOD is the used model (with or without main effect and with or without epistasis effect), OR is the odds ratio and PREV is the prevalence of the disease.