*Proceedings of the 2012 IEEE 12th International Conference on Bioinformatics & Bioengineering (BIBE), Larnaca, Cyprus, 11-13 November 2012*

# Clustering Subjects in Genetic Studies with Self Organizing Maps

Aristos Aristodimou
Department of Computer Science
University of Cyprus
Nicosia, Cyprus
Email: aristodimou.aristos@ucy.ac.cy

Athos Antoniades
Department of Computer Science
University of Cyprus
Nicosia, Cyprus
Email: athos@cs.ucy.ac.cy

Constantinos Pattichis
Senior Member, IEEE
Department of Computer Science
University of Cyprus
Nicosia, Cyprus
Email: pattichi@ucy.ac.cy

***Abstract*—Several machine learning techniques have been applied for finding multi-loci associations among Single Nucleotide Polymorphisms (SNPs) and a disease. In this paper it is investigated whether Self Organizing Maps (SOMs) can generate clusters associated with a disease based on the genetic patterns of subjects. A batch categorical SOM that can handle missing data was used on Genome Wide Association (GWA) data on Multiple Sclerosis (MS). The association of the clusters generated with the disease were initially tested using the Pearson's chi square test and then the weights of the top clusters were used for investigating for SNP patterns. The results of the analyses reveal statistically significant associations between the generated clusters and the disease, indicating that SOMs can be used for multi-loci associations.**

***Index Terms*—Self Organizing Map, Clustering, GWA, SNP, Multi-loci Association Testing.**

## I. Introduction

In common complex diseases, multi-loci interactions are more important than the main effect of any single SNP. Single locus association studies in such diseases may not replicate their results across multiple samples, due to the effect of epistasis and other phenomena [1]. In this paper, Self Organizing Maps (SOMs) are investigated for clustering Genome Wide Association (GWA) data for multi-loci association testing.

Traditional genetic analyses focus on single locus associations and not on multi-loci associations. Several machine learning techniques have been applied for multi-loci association testing [2]–[4]. Most of them are using Neural Networks (NNs), Support Vector Machines (SVMs), Random Forests (RFs), Multifactor Dimensionality Reduction (MDR) and variations of these techniques and they will be introduced.

Ritchie *et al* applied Multifactor Dimensionality Reduction (MDR) to a sporadic breast cancer data set [5], whereas variations of MDR were also introduced for multi-loci associations [6]–[8]. MDR and its variations had good results when they were used on a small number of SNPs but they cannot be directly used on a large number of SNPs, due to the exhaustive search it performs for identifying n-SNP associations [4].

NNs have also been applied in such studies [9], but their results are affected by the architecture of the network. Since it is computationally intractable to perform an exhaustive search for selecting the appropriate architecture, techniques based on genetic programming GPNN [10] and grammatical evolution GENN [11] were proposed with promising results. When these two optimisation methods were compared, GENN outperformed GPNN [11].

Wadell *et al* [12] used SVM to test their hypothesis that different SNP patterns exist among patients with Multiple Myeloma diagnosed at a young age and patients diagnosed after the age of 70. They obtained an accuracy of 71% in classifying these two patient classes giving them evidence that their initial hypothesis was correct. In [13] an SVM approach was proposed for gene-gene interaction with comparable results with MDR. The approach handled unbalanced data better than MDR and was less susceptible to overfitting, but it was computationally expensive [13]. A disadvantage of SVMs, is that they do not cope well with missing data [13].

RFs have also been used for multi-loci association testing [14], [15]. A variation of RF is SNPInterForest [16], which copes with some limitations of RFs such as the fact that they may ignore SNPs with low marginal effects and the difficulty of extracting the interactions patterns. SNPInterForest was applied on 10,000 SNPs and identified two novel interactions but the analysis was computationally demanding. Another variation of RFs is Random Jungle (RJ) [17], which is an efficient method for analysing GWA data. The method was applied on 275,153 SNPs revealing new interactions and validating findings of recent studies. As with RFs, RJ's findings were affected when SNPs only had weak main effecs [4].

SOMs [18] were proposed by Tuevo Kohonen and have been widely used for clustering data in several scientific areas. To the authors knowledge, they have not been used for identifying multi-loci associations, but they have been used on biological data analyses [19]–[22]. Classical SOMs were intended for use with numerical data, but SNPs are nominal categorical data, hence a SOM for categorical data needs to be used for such data. NCSOM [23] is an algorithm that was proposed for handling numerical and categorical variables with promising results, and its update rule for nominal categorical variables is also used in this paper.

In this study, the NCSOM algorithm was tested on GWA data, and specifically on subjects with and without MS. Initially a subset of the SNPs was selected and then the algorithm was trained with 10-fold cross validation for selecting the

*978-1-4673-4358-9/12/$31.00 ©2012 IEEE*

appropriate map size. Once the map size was defined, the algorithm was applied on the selected SNPs and its clustering results were evaluated.

The structure of the paper is as follows. Section II provides the methodology followed for the experiments carried out and in Section III the results of the experiments are presented. In Section IV the results are discussed and finally in Section V the concluding remarks are provided.

## II. Methodology

### A. Dataset

The data used are from Australia and New Zealand Multiple Sclerosis Genetics Consortium (ANZgene) [24]. The dataset has 1,618 people with MS (cases) and 3,413 people without MS (controls). The genotyping platform used was made by Illumina 300k model. The distribution of the subjects in the dataset is shown in the first three columns in Table I.

TABLE I
ANZgene Subjects Distribution

| | Males | Females | Total | Training | Testing |
|---|---|---|---|---|---|
| Cases | 445 | 1173 | 1618 | 1456 | 162 |
| Controls | 757 | 1231 | 1988 | 1789 | 199 |
| Total | 1202 | 2404 | 3606 | 3245 | 361 |

Before using the data for training SOM, they were encoded using the encoding schema shown in Table II. An allele value of "a" represents the minor allele of a SNP and "A" represents the major allele of that SNP. With this encoding, if a subject has homozygous minor allele in a SNP and another subject has homozygous major allele on that SNP the distance between the two subjects for that SNP will be two. In the case that one of the subjects has a heterozygous allele and the other has a homozygous allele the distance will be one. The missing alleles were encoded with a -1 so that SOM can identify them, since it handles missing data differently.

TABLE II
Data Encoding

| Allelle 1 | Allele 2 | Encoding Value |
|---|---|---|
| a | a | 0 1 |
| a<br>A | A<br>a | 1 1 |
| A | A | 1 0 |
| Missing | Missing | -1 -1 |

### B. Feature Selection

The dataset consists of approximately 300,000 SNPs per subject, but this paper focuses on SNPs in the HLA region, where previous studies have identified associations among the region and MS [25]–[28]. A subset of SNPs in the regions was selected using a two SNP interaction algorithm [29], [30]. Specifically, the SNP pairs that were found to be associated with the disease by the two SNP interaction algorithm were selected as input for the training of SOM. This analysis resulted in a total of 37 SNPs. The advantage of using a two SNP interaction algorithm for feature selection with SOM, is that we can test for associations among the selected SNPs. For example if the two SNP interaction algorithm returned that SNP A and SNP B were associated and that SNP A and SNP C were associated, SOM would also cluster the association of SNP A, B and C with the disease.

### C. Categorical SOM for clustering SNPs

A batch categorical SOM was used in this paper, that uses the update rule for nominal categorical data from [23] with some modifications for handling missing data. Specifically, missing data were "ignored", since they were always considered as a match when compared with any allelic value. SOM consisted of an input layer which selected an input vector at a time as the input of the network, and an output layer that had the neurons that represented the final clusters. The input layer was an $N$ dimensional vector $\mathbf{x}$, where $N$ was the number of features of the $P$ input vectors. The output layer consisted of the neurons mapped in a two dimensional map. The number of neurons and the size of the map was predefined before training. Each neuron consisted of an $N$ dimensional vector $\mathbf{w}$ called the weights of the cluster and the $i$th weight of each cluster corresponded to the $i$th feature of the input vector. The set $a = \{0, 1, -1\}$ represents the possible categorical values of each input vector feature and $a_r$ represents the categorical value with index $r$, where $r = 1, 2, 3$.

**Algorithm 1** Categorical SOM

Initialize the weights ($\mathbf{w}$) with random values $\{0,1\}$
$t = 1$ (current epoch)
**Require:** $radius$ (neighbourhood radius), $T$ (final epoch)
**while** $t \neq T$ **do**
  **for each** input vector $x^j$ **do**
    find its BMU using (1)
  **end for**
  update the weights of each neuron using (4)
  $t = t + 1$
  reduce $radius$
**end while**

The training of the network is shown in Algorithm 1. Initially the weights of the neurons were randomly set to zeros and ones. Then the best matching unit (BMU) of each input vector was calculated using (1). As can be seen in (3) any feature with the missing value was considered as a matching case with any weight value.

$$BMU^j = argmin_m D(x^j, w^m) \tag{1}$$

where

$$D(x^j, w^m) = \sum_{i=1}^{i=N} \delta(x_i^j, w_i^m) \tag{2}$$

$$\delta(x_i^j, w_i^m) = \begin{cases} 0 & if(x_i^j = w_i^m) \\ 0 & if(x_i^j = -1) \\ 1 & if(x_i^j \neq w_i^m) \end{cases} \tag{3}$$

Once the BMU of each input vector was calculated the weights of each neuron were updated using (4). As seen in (4), the weights get the value of the most frequent categorical value of each feature (ignores the categorical value of missing data). The function $h_{BMU^j m}$ is a Gaussian neighbourhood function centred at the BMU of $x^j$.

$$w_i^m(t+1) = \begin{cases} a_c & if(F(a_c, w_i^m(t)) > F(a_{r\notin\{c,3\}}, w_i^m(t))) \\ a_c & if(F(a_c, w_i^m(t)) = F(a_{r\notin\{c,3\}}, w_i^m(t)) \\ & \wedge\ random(0,1) > 0.5) \\ w_i^m(t) & otherwise \end{cases} \tag{4}$$

where

$$F(a_r, w_i^m) = \frac{\sum_{j=1}^{j=P} (h_{BMU^j m} | x_i^j = a_r \vee x_i^j = -1)}{\sum_{j=1}^{j=P} h_{BMU^j m}},\ r = 1, 2 \tag{5}$$

$$c = arg_r max F(a_r, w_i^m),\ r = 1, 2 \tag{6}$$

Once the weights were updated, the neighbourhood ($radius$) was decreased. Then the process was repeated for a predefined number of epochs $T$.

### D. Models Investigated

Different map sizes were evaluated on the input dataset using a 10-fold cross validation. Each fold had the same proportion of cases and controls based on the initial distribution of cases and controls of the dataset. Specifically each fold contained approximately 162 cases and 199 controls and the total number of subjects used in training and testing are shown in Table I. Due to the initial randomization of the weights of SOM, the 10-fold cross validation procedure was repeated 10 times. To select the appropriate map size for SOM, the test was repeated on a 2x2, a 4x4, a 6x6 and an 8x8 map size for 1000 epochs. The initial neighbourhood radius was set as half of the map's edge size.

### E. Evaluation Metrics

*1) Traditional Measures:* To evaluate the classification power of SOM and test how the map size of the network affects the results, each cluster of the trained network was labeled as "case" if the majority of its subjects were cases and "control" otherwise.

TABLE III
CONFUSION MATRIX

| | | Actual | |
|---|---|---|---|
| | | Control | Case |
| Predicted | Control | TP | FP |
| | Case | FN | TN |

Then the confusion matrix was calculated as shown in Table III and the following evaluation metrics were calculated:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{8}$$

$$Specificity = \frac{TN}{TN + FP} \tag{9}$$

$$BalancedAccuracy = \frac{Sensitivity + Specificity}{2} \tag{10}$$

*2) Pearson's Chi Square Test ($\chi^2$):* Because the evaluation metrics mentioned above do not take into consideration the map size, $\chi^2$ was also used for evaluating the clustering of the network. The hypothesis here is that the generated clusters will be associated with the case/control status. The $\chi^2$ provides a well established methodology to test that. Specifically the null hypothesis tested, is that there is no association among the clusters generated and the case/control status of the subjects in the clusters.

For testing this, two contingency tables were created after the training of each SOM run. The first had the number of cases and controls of each neuron (cluster) using the training data and the second used the testing data. Then the p-value of $\chi^2$ was calculated on each contingency table. It was decided a-priori that the null hypothesis would be rejected if the p-value was smaller than 0.01 ($-log(p\text{-}value) > 2$). The advantage of this test is the fact that the p-value calculated takes into consideration not only the distribution of the cases and controls in each cluster, but also the number of active clusters. An active cluster in this paper is defined as a cluster with at least one subject assigned in it.

### F. Pattern Identification

After training, the best performing model was selected. The parameters of this model (e.g. map size, neighbourhood radius) were used for constructing an SOM, where the whole dataset was used for training. Once the model was trained, the generated clusters were tested for association with the case-control status of the subjects using $\chi^2$. The $\chi^2$ was calculated using a contingency table with the number of cases and controls of each cluster. If the statistical significance of the association was above the predefined threshold ($-log(p\text{-}value) \geq 2$), the patterns of the clusters would be further investigated to identify SNP associations with the disease status of the subjects clustered. Since each weight of a cluster represents the allelic value of the majority of the subjects of that cluster, the weights of the trained SOM were used to identify any interesting patterns among SNPs. To have more representative results (since the weight values were dependent on the distribution of the two classes in a cluster), only clusters with a minimum separation of 70% - 30% among the two classes were selected for pattern identification.

## III. RESULTS

### A. Models Investigated

A total of 37 SNPs were selected by the two SNP interaction algorithm [29] and these SNPs were used as input for all tests

TABLE IV
CLUSTERING RESULTS USING THE TOP 37 SNPS IN THE HLA REGION

| | | SOM map size | | | |
|---|---|---|---|---|---|
| | | 2x2 | 4x4 | 6x6 | 8x8 |
| $\chi^2$ $(-log(p\text{-}value))$ | Train | 21.0±2.1 | **42.9**±2.9 | 39.9±3.0 | 35.1±2.5 |
| | Test | 2.7±1.4 | **3.1**±1.4 | 2.0±1.0 | 1.7±0.4 |
| Accuracy (%) | Train | 58±0.3 | 63±0.7 | 63±0.7 | **64**±0.8 |
| | Test | 58±3 | 62±2.7 | 63±2.3 | **63**±2.2 |
| Sensitivity (%) | Train | 64±2.0 | 66±3.5 | 70±1.9 | **70**±1.7 |
| | Test | 64±4 | 66±4.2 | **69**±2.9 | 69±3 |
| Specificity (%) | Train | 51±2.7 | **58**±4 | 56±2.6 | 57±2.5 |
| | Test | 51±5 | **58**±6.4 | 55±5.4 | 55±5.4 |
| Balanced Accuracy (%) | Train | 58±0.4 | 62±0.8 | 63±0.8 | **64**±0.8 |
| | Test | 57±2.6 | 62±2.8 | 62±2.5 | **62**±2.4 |

performed. The results of the 10-fold cross validation for both the train and test sets are presented in Table IV. From the $\chi^2$ test, it is clear that the best map size for the selected dataset is the 4x4 map size. Specifically there is a major increase in the $-log(p\text{-}value)$ from the 2x2 size to the 4x4 size and then it decreases as the map size increases. The $-log(p\text{-}value)$ was $42.9 \pm 2.9$ for the training set and $3.1 \pm 1.4$ for the testing set, which are above the predefined threshold. For the testing set, the measure was close to the predefined threshold, but this was mainly because of the small number of subjects used in the testing phase. To address this, models were retrained using half of the subjects for training and the other half for testing on 5 different such sets. SOM was run 10 times on each set using the 4x4 map size and statistically significant clusters were obtained for both training and testing results with a $-log(p\text{-}value)$ close to 21, which is far greater than the predefined threshold.

The percentage of balanced accuracy was $62 \pm 2.8$ for the test set of the 4x4 map size. Similar to this value but with slightly smaller standard deviations were obtained for the 6x6 and 8x8 map sizes. The percentage of sensitivity and specificity were $66 \pm 3.5$ and $58 \pm 6.4$ for the test set for the 4x4 map size. Similar values were also obtained for the 6x6 and 8x8 map sizes.

### B. Pattern Identification

After training the SOM with a 4x4 size map using all of the subjects and the top 37 SNPs, the $-log(p\text{-}value)$ was 45, indicating that the weights of its clusters were associated with the disease. In this model, 5 major clusters were identified with more than 65% separation among the normalized distribution of cases and controls. Two of them are "case" clusters and the others are "control" clusters. In total these 5 clusters account for 1312 subjects, which is approximately 36% of the total number of subjects. These 5 clusters are highlighted with a grey background in Fig. 1. In each pie chart there are two numbers where the number in the blue area represents the fraction out of the total controls that got clustered in that specific cluster and similarly the same applies for the number in the red area and cases.

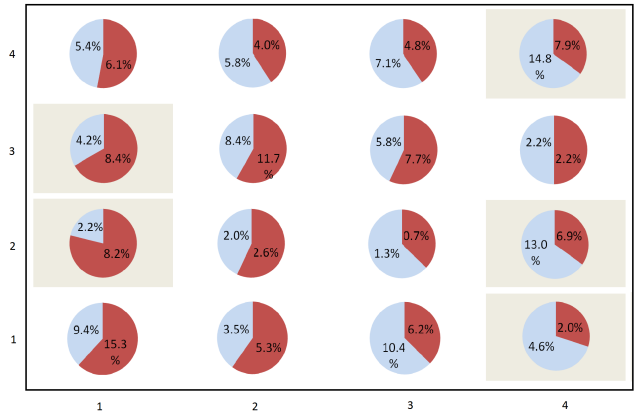


Fig. 1. The normalized distribution of cases and controls in each cluster after training SOM on a 4x4 map using all of the subjects for training with the top 37 SNPs. Light blue represents the controls and red represents the cases. The values in the blue areas in the pie charts represent the proportion out of the total controls that are in each cluster and similarly the numbers in the red areas the proportions out of the total cases. Highlighted with a grey background are the top 5 clusters based on the difference of their distribution of cases and controls.

Out of these 5 clusters, 4 of them had a proportion of cases and controls above the a-priori defined threshold (70% - 30%). In Fig. 2 each row illustrates the weights of these 4 clusters. The numbers in the first column represent the (x,y) coordinates of each cluster as illustrated in Fig. 1. The second column has bar charts showing the distribution and the number of cases and controls in each cluster. The numbers in parentheses in the bar charts, represent the fraction out of the total number of cases and controls for each cluster. The rest of the columns represent the weight values of the two alleles for each SNP. As can be seen, the "control" clusters have similarities among them that are not present in the "case" cluster. Three SNPs that illustrate this are RS1053924, RS926070 and RS3129941, which are highlighted with red rectangles. In these three SNPs,

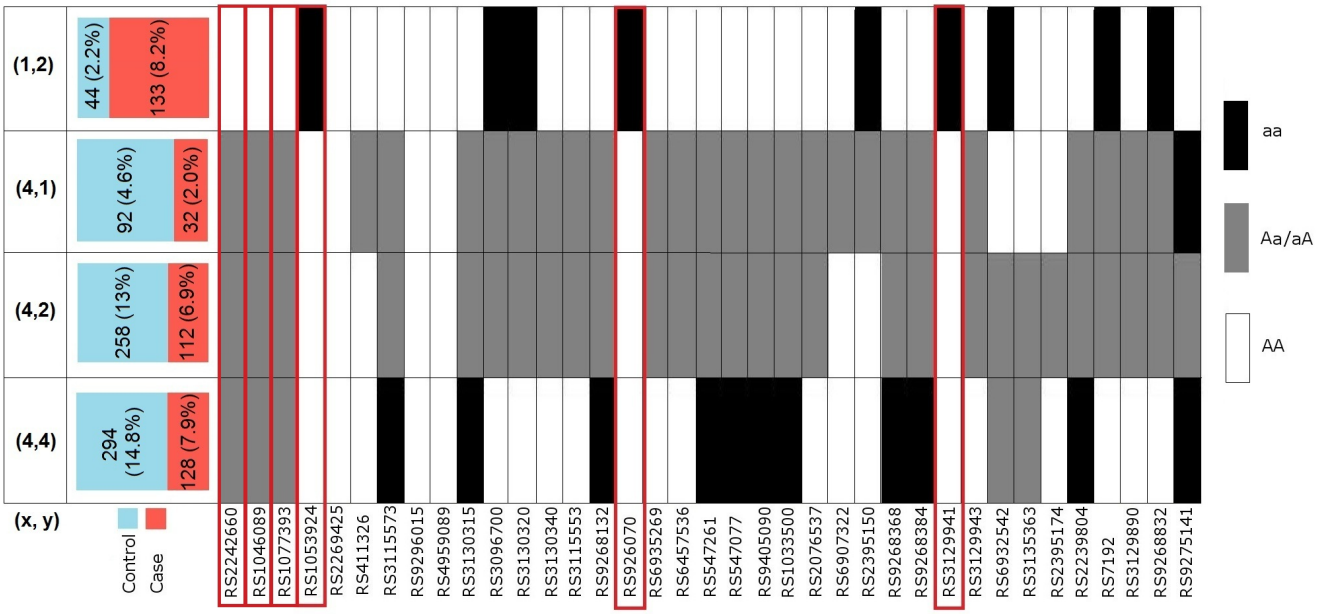


Fig. 2. Pattern identification showing the top 4 clusters of a 4x4 map using the top 37 SNPs as input. The numbers in the first column represent the (x,y) coordinates of each cluster as illustrated in Fig. 1. The second column has bar charts with details such as the number of cases and controls in each cluster and the fraction out of the total number of cases and controls that each number corresponds in parentheses. The rest of the columns have the weight values of the clusters, that represent the most frequent allele values of the subjects of that cluster for the specific SNP. "AA" represents homozygous major allele value, "aa" homozygous minor allele value and "aA/Aa" heterozygous allele value for a SNP. The highlighted columns (RS2242660, RS1077393, RS1046089, RS1053924, RS926070, RS3129941) show important differences in patterns among the case and controls clusters.

the majority of the subjects in the case cluster had homozygous minor allele values whereas in the controls clusters, the majority had homozygous major allele values. Similarly for the first three SNPs (RS2242660, RS1077393, RS1046089) in Fig. 2, the case cluster had homozygous major allele values, whereas the control clusters had heterozygous allele values. The two SNP interaction algorithm indicated for these three SNPs that RS2242660 interacted with RS1077393 and that RS1077393 interacted with RS1046089, hence SOM identified a three SNP interaction among these SNPs.

## IV. Discussion

In this paper SOM, a clustering algorithm, was tested to investigate whether it could find clusters whose SNPs were strongly associated with MS. From the results obtained, an important finding is that the clusters identified are statistically significant. The top clusters had an important separation among cases and controls and they accounted for a good proportion of the total subjects of the dataset. Moreover, some interesting patterns among the top case and controls clusters were identified. Further investigation of these patterns needs to be performed for identifying the actual causative effects driving them.

Many studies indicated that there is an association among the HLA region and MS using single locus association testing [25]–[28]. Antoniades used a two SNP interaction algorithm in [29] and was able to identify statistically significant two SNP interactions among 37 SNPs, which were replicated using an independent dataset. Those 37 SNPs were used in this paper and SOM was able to identify higher order SNP interactions, with 6 SNPs revealing interesting patterns among the "case" and "control" clusters. Brassat *et al* identified single and three locus association models among SNPs in the HLA region and MS using MDR [31], whereas in [32], Motsinger *et al* identified two, three and four locus associations among SNPs in the HLA region and MS using MDR as well. The SNPs identified by these two studies were not selected by the feature selection algorithm used in this paper, hence the associations identified in this paper cannot be compared with the associations identified in those two studies. The findings of this paper indicate that SOMs can be used for clustering GWA data for finding associations among SNPs and a disease. The advantage of the SOM is that it is searching for n-SNP associations when creating the clusters. This is accomplished by clustering subjects together that have as many similar SNP patterns as possible without performing an exhaustive search as MDRs [4].

From the metrics used for evaluating the clustering and selecting the map size of the network, the $\chi^2$ metric was more indicative than the traditional evaluation metrics. It has the advantage of considering the initial distribution of the classes, which is important when analysing imbalanced data. This is something that the standard accuracy measure does not cope with, making it inappropriate for such cases. Balanced accuracy can be used with imbalanced data and when used with specificity and sensitivity, the accuracy on each class and their average accuracy can be observed. But these measures

do not take into consideration the number of clusters used by the network as the calculation of the p-value of $\chi^2$ does. Moreover the p-value calculated accounts for the number of subjects used as well, hence $\chi^2$ can be used as a single measure for evaluating and selecting the map size of the network.

## V. Conclusion

The ability of SOM for finding associations among SNPs and a disease has been investigated with promising results. From the results obtained, it can be seen that the unsupervised clustering of SOM was statistically significant, revealing an association among the SNP patterns in the clusters generated and the disease status of the subjects. Moreover the $\chi^2$ statistic due to its ability of taking into consideration the distribution of the classes, the number of subjects in the dataset and the number of active clusters of the network, has been proposed for selecting the map size of the network instead of the traditional evaluation measures. Finally we conclude that SOMs ability of clustering subjects with similar SNP patterns together, is an important feature that may prove of significant value in future multi-loci association testing.

As part of future work, independent datasets will be used for replication testing and further investigation of the association of the patterns found and MS will be performed. Furthermore, larger and smaller subsets of the SNPs will be used to test how the inclusion and exclusion of SNPs affects the results of the algorithm. For testing how the rate of missing data affects the clustering of the proposed algorithm, different rates of missing values will be included in the dataset used in this paper.

## References


[1] J. Moore, "The ubiquitous nature of epistasis in determining susceptibility to common human diseases," *Human heredity*, vol. 56, no. 1-3, pp. 73–82, 2003.

[2] B. A. McKinney *et al.*, "Machine learning for detecting gene–gene interactions: a review," *Appl. bioinformatics*, vol. 5, no. 2, pp. 77–88, 2006.

[3] H. J. Cordell, "Detecting gene–gene interactions that underlie human diseases," *Nature Reviews Genetics*, vol. 10, no. 6, pp. 392–404, Jun. 2009.

[4] R. Upstill-Goddard *et al.*, "Machine learning approaches for the discovery of gene-gene interactions in disease data," *Briefings in Bioinformatics*, May 2012.

[5] M. D. Ritchie *et al.*, "Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer," *The Amer. J. of Human Genetics*, vol. 69, no. 1, pp. 138–147, Jul. 2001.

[6] Y. Chung *et al.*, "Odds ratio based multifactor-dimensionality reduction method for detecting gene–gene interactions," *Bioinformatics*, vol. 23, no. 1, pp. 71–76, Jan. 2007.

[7] X.-Y. Lou *et al.*, "A generalized combinatorial approach for detecting gene-by-gene and gene-by-environment interactions with application to nicotine dependence," *Amer. J. of Human Genetics*, vol. 80, no. 6, pp. 1125–1137, Jun. 2007.

[8] J. Gui *et al.*, "A robust multifactor dimensionality reduction method for detecting Gene–Gene interactions with application to the genetic analysis of bladder cancer susceptibility," *Ann. of Human Genetics*, vol. 75, no. 1, p. 2028, 2011.

[9] D. Curtis, B. V. North, and P. C. Sham, "Use of an artificial neural network to detect association between a disease and multiple marker genotypes," *Ann. of Human Genetics*, vol. 65, no. 1, p. 95107, 2001.

[10] A. A. Motsinger *et al.*, "GPNN: power studies and applications of a neural network method for detecting gene–gene interactions in studies of human disease," *BMC bioinformatics*, vol. 7, p. 39, 2006.

[11] A. A. Motsinger-Reif *et al.*, "Comparison of approaches for machine-learning optimization of neural networks for detecting gene–gene interactions in genetic epidemiology," *Genetic epidemiology*, vol. 32, no. 4, pp. 325–340, May 2008.

[12] M. Waddell, D. Page, and J. Shaughnessy, Jr., "Predicting cancer susceptibility from single-nucleotide polymorphism data: a case study in multiple myeloma," in *Proc. of the 5th Int. Workshop on Bioinformatics*, ser. BIOKDD '05. New York, NY, USA: ACM, 2005, pp. 21–28.

[13] S.-H. Chen *et al.*, "A support vector machine approach for detecting gene–gene interaction," *Genetic Epidemiology*, vol. 32, no. 2, p. 152167, 2008.

[14] A. Bureau *et al.*, "Identifying snps predictive of phenotype using random forests," *Genetic Epidemiology*, vol. 28, no. 2, pp. 171–182, 2005.

[15] K. Lunetta *et al.*, "Screening large-scale association study data: exploiting interactions using random forests," *BMC genetics*, vol. 5, no. 1, p. 32, 2004.

[16] M. Yoshida and A. Koike, "SNPInterForest: a new method for detecting epistatic interactions," *BMC Bioinformatics*, vol. 12, no. 1, p. 469, Dec. 2011.

[17] D. F. Schwarz, I. R. Knig, and A. Ziegler, "On safari to Random Jungle: a fast implementation of Random Forests for high-dimensional data," *Bioinformatics*, vol. 26, no. 14, pp. 1752–1758, Jul. 2010.

[18] T. Kohonen, "The self-organizing map," *Proc. of the IEEE*, vol. 78, no. 9, pp. 1464 –1480, Sep. 1990.

[19] P. Tamayo *et al.*, "Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation," *Proc. of the Nat. Academy of Sci.*, vol. 96, no. 6, pp. 2907–2912, Mar. 1999.

[20] C. Martin *et al.*, "Hyperbolic SOM-based clustering of DNA fragment features for taxonomic visualization and classification," *Bioinformatics*, vol. 24, no. 14, pp. 1568–1574, Jul. 2008.

[21] A. M. Newman and J. B. Cooper, "AutoSOME: a clustering method for identifying gene expression modules without prior knowledge of cluster number," *BMC Bioinformatics*, vol. 11, no. 1, p. 117, Mar. 2010.

[22] G. Skreti, E. Bei, and M. Zervakis, "Shape-influenced clustering of dynamic patterns of gene profiles," in *Int. Conf. of the IEEE Eng. in Med. and Biol. Soc.*, San Diego, California, Sep. 2012.

[23] N. Chen and N. Marques, "An extension of self-organizing maps to categorical data," in *Progress in Artificial Intell.*, ser. Lecture Notes in Computer Science, C. Bento, A. Cardoso, and G. Dias, Eds. Springer Berlin / Heidelberg, 2005, vol. 3808, pp. 304–313.

[24] M. Bahlo *et al.*, "Genome-wide association study identifies new multiple sclerosis susceptibility loci on chromosomes 12 and 20," *Nature genetics*, vol. 41, no. 7, pp. 824–828, Jul. 2009.

[25] J. R. Oksenberg *et al.*, "The genetics of multiple sclerosis: SNPs to pathways to pathogenesis," *Nature Reviews Genetics*, vol. 9, no. 7, pp. 516–526, Jun. 2008.

[26] G. Ren *et al.*, "Association of the HLA region with multiple sclerosis as confirmed by a genome screen using >10,000 SNPs on dna chips," *J. of Molecular Medicine*, vol. 83, no. 6, pp. 486–494, 2005. [Online]. Available: http://www.springerlink.com/content/jwx8864177q6v524/abstract/

[27] J. Link *et al.*, "Two HLA class I genes independently associated with multiple sclerosis," *J. of Neuroimmunology*, vol. 226, no. 12, pp. 172–176, Sep. 2010.

[28] D. Hafler *et al.*, "Risk alleles for multiple sclerosis identified by a genomewide study," *The New England J. of medicine*, vol. 357, no. 9, pp. 851–862, Aug. 2007.

[29] A. Antoniades, "Discovering disease associated gene–gene interactions: A two snp interaction analysis framework," *Ph.D. dissertation, University of Cyprus, Cyprus*, 2011.

[30] A. Antoniades *et al.*, "A computationally fast measure of epistasis for 2 SNPs and a categorical phenotype," *IEEE EMBC*, vol. 2010, pp. 6194–6197, 2010.

[31] D. Brassat *et al.*, "Multifactor dimensionality reduction reveals gene–gene interactions associated with multiple sclerosis susceptibility in African Americans," *Genes and Immunity*, vol. 7, no. 4, pp. 310–315, 2006.

[32] A. Motsinger *et al.*, "Complex gene–gene interactions in multiple sclerosis: a multifactorial approach reveals associations with inflammatory genes," *Neurogenetics*, vol. 8, no. 1, pp. 11–20, 2007.