2006 International Joint Conference on Neural Networks
Sheraton Vancouver Wall Centre Hotel, Vancouver, BC, Canada
July 16-21, 2006

A Neural Network Model for Maximizing Prediction Accuracy in Haplotype Tagging SNP Selection

Jae-Yoon Jung* and Phil Hyoun Lee†
*Department of Computer Science, University of Maryland
College Park, MD 20742 USA
E-mail: jung@cs.umd.edu
†School of Computing, Queen's University
Kingston, ON K7L 3N6 Canada
E-mail: lee@cs.queensu.ca

***Abstract*— Due to the tremendous number of single nucleotide polymorphisms (SNPs), there is a clear need to expedite genotyping by considering only a subset of all SNPs called haplotype tagging SNPs (htSNPs). Recently, the approach that selects htSNPs by maximizing their prediction accuracy has demonstrated very promising results. Here we propose a new prediction system for htSNP selection based on neural network models. We applied our system to three public data sets, and compared its prediction performance to that of two state-of-the-art prediction rules. The results demonstrate that our system consistently outperforms compared methods with robust performance.**

I. Introduction

A large number of single nucleotide polymorphisms (SNPs) in the human genome have greatly expedited the progress of biomedical research. The high-density genetic map of SNPs enables close examination of a disease locus, and as a result, plays a fundamental role in large-scale association studies to locate common and complex disease genes [1]. However, due to the sheer number of SNPs, which is estimated at about ten million [2], genotyping all of them in a candidate region is still costly and time-consuming. Therefore, selecting a subset of SNPs that is sufficiently informative to conduct disease-gene association studies but still small enough to reduce the genotyping overhead, is a major challenge. The selected SNPs are generally referred to as *haplotype tagging SNPs* (htSNPs) [3], and the unselected SNPs are referred to as *tagged SNPs*. In short, htSNP selection problem refers to find an optimal set of htSNPs to predict the other SNPs.

Numerous computational methods have been developed for conducting htSNP selection. Their common goal is to identify a subset of SNPs that can distinguish all the common haplotypes [4], or at least explain a certain percentage of the haplotype diversity [3], [5], [6], [7]. They are all based on the assumption that even when the causal variant is not one of the selected htSNPs, those selected would still capture the haplotype susceptibility to the disease. Early studies typically concentrated on a small number of SNPs, and relied on exhaustive search [3], [4], [5], [6], [7]. More recent studies avoid exhaustive search by using clustering methods based on pairwise linkage disequilibrium (LD, non-random association of SNPs [8]) [9], matrix decomposition methods [10], greedy methods [11], or branch-and-bound algorithms [12].

Recently, several researchers have proposed a different approach, selecting htSNPs based on how well they predict the remaining set of tagged SNPs. Bafna et al. [13] and Halldorsson et al. [14] suggested a new measure called *informativeness*, which quantifies the confidence by which one group of SNPs can predict another. Halperin et al. [15] also proposed a new measure directly evaluating the prediction accuracy of a set of SNPs. By limiting the number of predictive SNPs or restricting them to a w-bounded neighborhood (where w is a fixed window size ≤ 30), both methods could identify the minimal set of htSNPs based on their respective measure. On the other hand, Lin et al. [16] selected htSNPs by using principal component analysis (PCA), and predicted the genotype of a tagged SNP using the one htSNP whose correlation coefficient with the tagged is the highest.

In general, after the selected htSNPs are genotyped, the alleles of the tagged SNPs are predicted using the alleles of the htSNPs, and disease-gene association is conducted based on the reconstructed full haplotype data. Therefore, above tagged SNP prediction methods present a prediction rule for tagged SNPs along with a selected set of htSNPs. Currently, all of them rely on a simple majority vote or pairwise-correlation to predict the alleles (i.e., the nucleotide at a position in which a SNP occurred) of tagged SNPs.

In this paper, we propose a new prediction system for htSNP selection problem. We use a neural network to learn the relationship among SNPs, by training it to produce the same haplotype pattern with the input. In the prediction phase, we build an input haplotype pattern based on htSNPs and the training data, then the trained neural network predicts the corresponding haplotype. To date, there has been only a small number of neural network approaches used in the genetic problems, such as haplotype association [17], [18], [19] and haplotype phasing [20]. And to our best knowledge, our approach is the first neural network model to tackle the htSNP selection problem.

We applied our method to three public data sets. The results based on one-leave-out cross validation show that

0-7803-9490-9/06/$20.00/©2006 IEEE

For each $k \leq K$,
  Repeat R times,
    Build T_k that randomly selects k SNPs.
    For each testing haplotype h_t,
      Train the network N with the remaining $D_{tr} = D - h_t$.
      Build a candidate haplotype h_{in} :
        For each haplotype pattern $h_i \in D_{tr}$,
          $c \leftarrow 0, \quad m \leftarrow 0$.
          For each SNP s_j^t, where $j \in T_k$,
            If s_j^i matches with s_j^t, **Then** $c \leftarrow c + 1$.
          If $c \neq 0$ **Then** $h_{in} \leftarrow h_{in} + c \cdot h_i, \ m \leftarrow m + 1$.
        $h_{in} \leftarrow h_{in} \ / \ m$.
      Get the output pattern $h_{out} \leftarrow N(h_{in})$ and calculate the prediction accuarcy from $(h_t - h_{out})$.
    Calcaulate the averaged prediction accuracy over all h_t.
Report T_{k^*} that maximizes the prediction accuracy.

Fig. 1. The algorithm for choosing htSNPs that maximizes the prediction accuracy. We assume that the maximum number of htSNPs K, and the size of random sampling R are given.

our prediction system achieves a higher level of prediction accuracy compared to that of a majority vote and pairwise correlation. Moreover, the performance of our system is robust under various selections of htSNPs or the number of htSNPs.

II. Prediction Method

Our approach to the prediction problem consists of two phases. In the learning phase, a neural network is used to discover the relationship between SNPs. In the prediction phase, first we try to build a candidate haplotype based on the given htSNPs. This prediction is then refined through the trained neural network. Figure 1 shows the overall algorithm, of which notations and procedures are explained in detail in the following sections.

A. Problem Formulation

Let us assume that we have a data set D consisting of n haplotypes $h_1, ..., h_n$ and each haplotype h_i has p SNPs $s_1^i, ..., s_p^i$. As we are interested in biallelic SNPs (i.e., only two types of variations are observed in a site), we consider a haplotype h of length p as a string of $\{-1, 1\}^p$, and D as an $n \cdot p$ matrix.

The prediction problem is to generate values (alleles) of SNPs using htSNPs. Assuming we are given a set of htSNPs T_k, which specifies k SNP positions (i.e., column indices in D) to look up, we try to rebuild the corresponding haplotype h_t, based on htSNP values s_j^i, where $i \neq t$ and $j \in T_k$. In other words, the prediction algorithm is a function of $f : \{-1, 1\}^k \rightarrow \{-1, 1\}^p$. In addition, we assume that the prediction of each SNP only depends on the values of htSNPs, not on the prediction results of the other SNPs. This assumption is required to predict tagged SNPs in arbitrary order.

B. Neural Network Model

We used a neural network model to learn the nonlinear higher-order relationship between SNPs. For each haplotype input pattern, the network is trained to generate the same haplotype as the given input. As shown in Figure 2, it is a fully-connected network of input size p, but the direct weights w_{ii} from the input to the hidden layer are removed in order to reduce the influence from the direct input nodes. We adopted the resilient error backpropagation (RPROP) [21] with the maximum weight change value of 1.0. In the prediction phase, we used one-leave-out cross validation [22] to calculate the averaged prediction accuracy. More specifically, we pick a haplotype h_t as a test data, and train our network with the remaining $n - 1$ haplotypes $D_{tr} = \{h_i \mid i \neq t\}$. The prediction accuracy is calculated by comparing the selected haplotype pattern with the output pattern of our network. This algorithm loops around for each haplotype, and the final prediction accuracy is averaged.

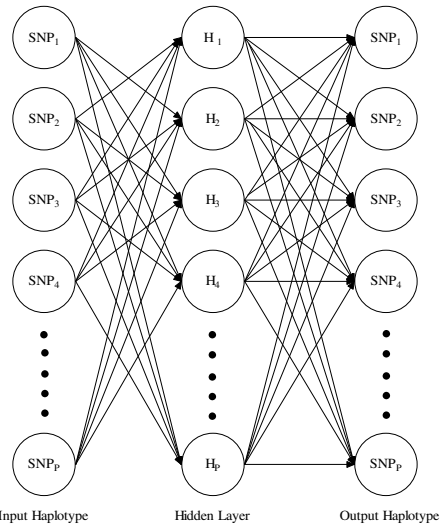

Fig. 2. The description of the neural network model for the prediction problem. Each node corresponds to a SNP. The direct weights w_{ii} from the input to the hidden layer are removed to facilitate learning dependency between different SNPs.

C. Weighted Vote

In the prediction phase, we make a candidate haplotype pattern h_{in} based on htSNPs T_k, then put this pattern into the network to predict the target (tagged) haplotype h_t. The algorithm to build the candidate haplotype pattern is as follows. For a given htSNP s_i^t where $i \in T_k$, we look up each haplotype $h_j \in D_{tr}$ to check if the SNP s_i^j matches up with s_i^t. If it matches, we add this haplotype h_j into the input pattern h_{in}. After looking up all haplotypes in the training data set, we average the current input by the number of the added haplotypes. In this way, we can implement the weighted majority vote for the candidate haplotype, which is proportional to the number of matched SNPs.

D. Random Sampling

Selecting the optimal set of tagging SNPs that maximizes prediction accuracy is NP-hard in general [13]. As we are focused in maximizing the prediction accuracy given a set of htSNPs, we use a simple algorithm to choose htSNPs when the number of htSNPs k is given. We build 100 sets of htSNPs, $T_1, T_2, ..., T_{100}$, where each set has the information of k randomly selected SNP sites. For each T_i, we calculate the accuracy and choose the T_{k^*} that maximizes the averaged accuracy over all testing haplotypes. We repeats this random sampling until k reaches 30 percents of the total SNPs in the given data set.

The same procedure is also applied to two compared methods: the correlation-based prediction method used in Lin et al. [16] and the majority-vote-based prediction method used in Bafna, Halldörsson et al. [13], [14], and Halperin et al. [15]. The correlation-based method predicts each tagged SNP using the single htSNP whose correlation coefficient with the tagged one is the highest. The majority-vote-based method consists of two sequential procedures; (1) the haplotypes in the training data, whose htSNP alleles are the same as those of the predicted haplotype, are identified. (2) each tagged SNP in the predicted haplotype is assigned the allele that occurs most often in the haplotypes identified. The majority-vote-based prediction is done with four different options: 1) using all htSNPs; 2) using only two nearest htSNPs as used in Halperin et al. [15]; 3) using only htSNPs in 13-bounded neighborhood as used in Bafna, Halldörsson et al. [13], [14]; and 4) using only htSNPs in 26-bounded neighborhood as used in Bafna, Halldörsson et al. [13], [14].

TABLE I

THE GENETIC CHARACTERISTICS OF THREE DATA SETS

Candidate Gene	ACE	LPL	IBD5
SNP No.	53	88	103
Haplotype No.	22	142	258
Gene Diversity	0.876	0.991	0.724
LD Mean (STD)	0.32 (0.34)	0.06 (0.156)	0.124 (0.174)

III. RESULTS

A. Test Data and Evaluation

Three public data sets, ACE (angiotensin I converting enzyme) [23], LPL (human lipoprotein lipase) [24], and IBD5 (inflammatory bowel disease 5) [25] are used for evaluation. Here we explain some genetic aspects of the data sets to facilitate the understanding of their characteristics. Gene diversity [26] measures the probability that two haplotypes chosen at random from the sample are different. Pairwise linkage disequilibrium (LD) between SNPs is estimated by LD correlation coefficient [8] (i.e., Δ^2). We use the χ^2 test with one degree of freedom in order to verify statistical significance of the standardized LD parameter. Table I summarizes the genetic characteristics of these three data sets.

Firstly, ACE contains 78 SNPs stretched over a genomic region of length 24 Kb on Chr. $17q23$. Among 78 SNPs, 52 bi-allelic non-singleton SNPs are analyzed. Genotyping is done for 11 individuals, and haplotype phasing (i.e., computational process of deducing haplotypes from genotypes) is done by PHASE [27]. The gene diversity is 0.876. Results of the four-gamete test on pairs of SNPs reveal that 19.38% pairs of loci show evidence for recombination or recurrent mutation. The average LD is 0.32 with a standard deviation 0.34.

Secondly, LPL contains 88 SNPs genotyped from 71 father-mother-child trios in the human lipoprotein lipase (LPL) gene. Haplotype information of this data is already known, and the estimate of the gene diversity is 0.99. Average LD is pretty low in this data set, since genotyping is done from three different populations.

Finally, the third data set IBD5 contains 103 SNPs on Chr. $5q31$ spanning 500 Kb with an average spacing of 1 to 133 Kb. The total of 129 father-mother-child trios from European-derived population is genotyped. Haplotype phasing is done by GERBIL [28], and the gene diversity of IBD5 is estimated as 0.98.

Each haplotype has a different occurrence frequency in a

TABLE II

P-VALUE SUMMARY OF T-TESTS ($\alpha = 0.01$)

Data	Compared Method	Neural Network Model
ACE	Correlation	6.7058e-016
	Majority vote	8.3855e-009
	Maj. 2 nearest	2.1681e-011
	Maj. 13 nearest	3.3901e-009
	Maj. 26 nearest	2.4763e-008
LPL	Correlation	3.6256e-020
	Majority vote	4.1656e-008
	Maj. 2 nearest	1.7744e-018
	Maj. 13 nearest	4.9208e-009
	Maj. 26 nearest	4.4456e-008
IBD5	Correlation	9.8103e-023
	Majority vote	7.1970e-008
	Maj. 2 nearest	4.3001e-014
	Maj. 13 nearest	0.2613
	Maj. 26 nearest	6.2719e-006

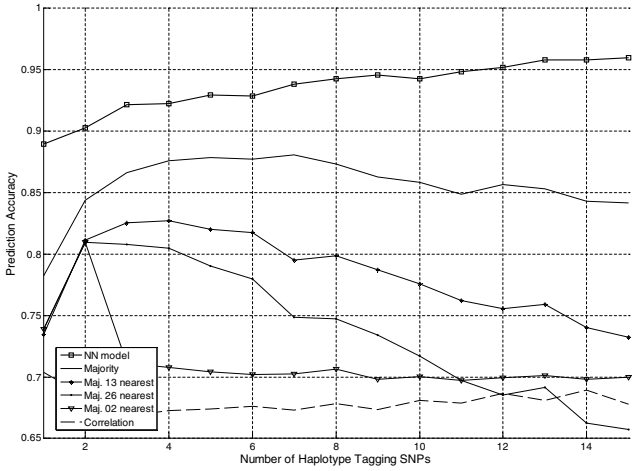

Fig. 3. Prediction accuracy results with ACE data set.

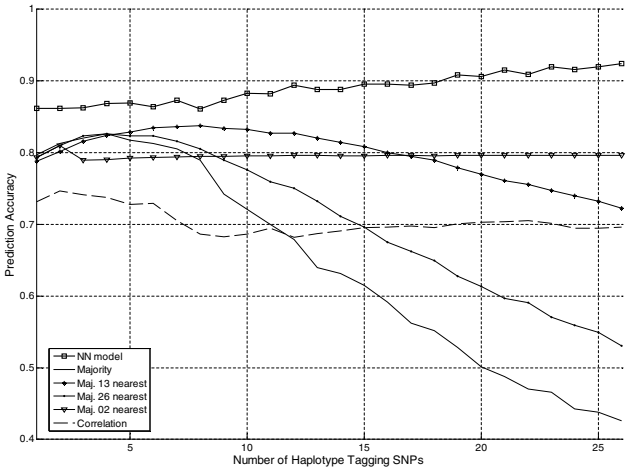

Fig. 4. Prediction accuracy results with LPL data set.

data set. For example, a specific haplotype in the IBD5 data set appears 134 times, while the other haplotypes typically occur just once or twice. In order to avoid bias during the learning phase, we build a training set with only unique haplotypes in a given data set, and later adjust their occurrence frequency when calculating the prediction accuracy.

B. Test Results

For the first two data sets, our approach clearly outperforms all the other compared methods in terms of the prediction accuracy, regardless of the number of htSNPs. These results are illustrated in Figure 3 and Figure 4. With the third data set IBD5, our approach shows similar performance with majority vote methods when the number of htSNPs are relatively small. However, it overcomes the other methods along with the number of htSNPs increasing, as shown in Figure 5. We performed the T-test to examine

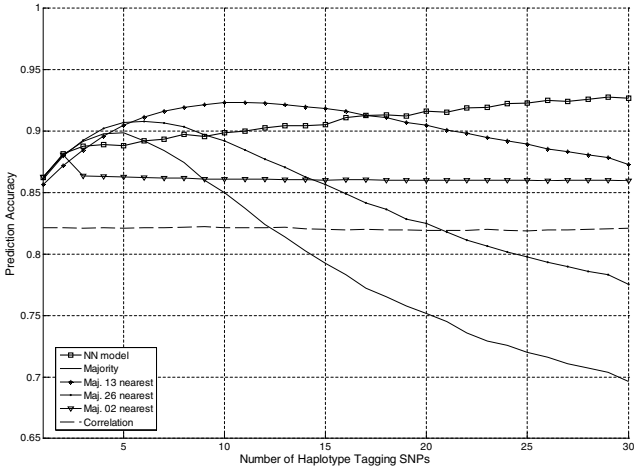

Fig. 5. Prediction accuracy results with IBD5 data set.

whether the difference between the performance of our method and that of the compared methods is significant. Using a significance level of one percent, all the performance differences except for one case with IBD5 data set were verified to be significant. Table II reports the p-values of these T-tests.

Moreover, the prediction accuracy of our approach shows positive correlation with the number of htSNPs in general. This type of robustness is an intuitively desired property of any prediction algorithm to be used in the htSNPs selection problem, as we expect to predict haplotypes more accurately when we are given more clues. Neither of compared methods seems to meet this requirement, as their performance rapidly decreases with increasing number of htSNPs (in majority vote based methods), or stays at the relatively low level (in the correlation based method).

IV. Discussion

Haplotype tagging SNP (htSNP) selection provides the most practical framework for conducting large-scale disease-gene association studies. In general, it can yield about 2-10 fold savings in the genotyping efforts with little loss of power in subsequent association studies [29], [30]. However, despite of its importance in current biomedical research, artificial neural networks have never been applied in the context of htSNP selection.

In this paper, we have presented a novel prediction system for htSNP selection based on a neural network model. The evaluation results demonstrate its superiority compared to two state-of-the-art prediction methods in terms of prediction accuracy and robustness.

As for future research directions, we plan to improve our system to efficiently handle: 1) rare haplotypes; 2) uncertainty in haplotype data; and 3) small sample size. Firstly, our prediction system performs well for common haplotypes or common SNPs but not for rare ones. Common variations are of interest because many common human diseases have

been explained by common DNA variations rather than by rare ones [31]. Furthermore, practically, a much larger sample size is needed to identify rare haplotypes [32]. However, it is still an open question whether common variations or rare ones influence the susceptibility to common and complex disease.

Second, currently we use haplotype data to predict the alleles of the tagged SNP. Thus, when only genotype data are available, haplotype phasing is performed on the genotype data, and the identified haplotype data are used as an input to our system. However, haplotype phasing may lead to incorrect resolution. To address this, some statistical algorithms produce multiple solutions along with their uncertainty, or the distribution of haplotype pairs for each genotype rather than a single resolved pair. We plan to incorporate this uncertainty of inferred haplotype data in our system.

Finally, only a small number of haplotypes are currently available for htSNP selection. Thus, to ensure that our prediction system built from a given haplotype sample works well for other samples from the same population, methods that can avoid over-fitting of the given data set should be considered as well.

What comprises the best htSNP selection strategy is still an open problem [29], and no standard evaluation measure has been established yet. Owing to its stable performance, we envision that our prediction system can be used as a common testbed for evaluating the performance of different selection approaches. Most of all, we consider our approach as a step further to challenging the htSNP selection problem.

REFERENCES

[1] B. S. Shastry, "Snps and haplotypes: Genetic markers for disease and drug response (review)," *International journal of molecular medicine*, vol. 11, 2003.

[2] S. T. Sherry, M. H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski, and K. Sirotkin, "dbSNP: the NCBI database of genetic variation," *Nucleic Acids Res.*, 29:308-311, 2001.

[3] G. C. L. Johnson, L. Esposito, B. J. Barratt, A. N. Smith, J. Heward, G. I. Genova, H. Ueda, I. A. Eaves, F. Dudbridge, R. C. J. Twells, F. Payne, W. Hughes, S. Nutland, H. Stevens, P. Carr, E. Tuomilegto-Wolf, J. Tuomilehto, S. C. L. Gough, D. G. Clayton, and J. A. Todd, "High-resolution haplotype strucutre in the human genome," *Nature Genetics*, vol. 29, no. 2, 2001.

[4] S. Gabriel, S. Scahffner, H. Nguyen, J. Moore, J. Roy, B. lumenstiel, J. Higgins, M. DeFelice, A. Lochner, M. Faggart, S. Liu-Cordero, C. Rotimi, A. Adeyemo, R. Cooper, R. Ward, E. Lander, M. Daly, and D. Altschuler, "The structure of haplotype blocks in the human genome," *Science*, vol. 296, 2002.

[5] X. Ke and L. R. Cardon, "Efficient selective screening of haplotype tag snps," *Bioinformatics*, vol. 19, no. 2, 2003.

[6] H. Ackerman, S. Usen, R. Mott, A. Richardson, F. Sisay-Joof, P. Katundu,T. Taylor, R. Ward, M. Molyneux, M. Pinder, and D. P. Kwiatkowski, "Haplotype analysis of the tnf locus by association efficiency and entropy," *Genome Biology*, vol. 4, no. 4, pp. R24.1–R24.13, 2003.

[7] H. I. Avi-Itzhak, X. Su, and F. M. D. L. Vega, "Selection of minimum subsets of single nucleotide polymorphism to capture haplotype block diversity," *In Proceedings of Pacific Symposium on Biocomputing*, 2003,

[8] X. Lu, T. Niu and J. S. Liu, "Haplotype information and linkage disequilibrium mapping for single nucleotide polymorphisms," *Genome Research*, 13:2112-2117,2003.

[9] S. I. Ao, K. Yip, M. Ng, D. Cheung, P. Fong, I. Melhado and P. C. Sham, "CLUSTAG: Hierarchical clustering and graph methods for selecting tag SNPs," *Bioinformatics*, 21:1735-1736,2005.

[10] Z. Meng, D. V. Zaykin, C. Xu, M. Wagner and M. G. Ehm, "Selection of Genetic Markers for Association analyses, using linkage disequilibrium and haplotypes," *Am. J. Hum. Genet.*, 73:115-130, 2003.

[11] C. S. Carlson, M. A. Eberle, M. J. Rieder, Q. Yi, L. Kruglyak, and D. A. Nickerson, "Selecting a Maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilirium," *Am. J. Hum. Genet.*, 74:106-120, 2004.

[12] K. Ding and J. Zhang and K. Zhou and Y. Shen and X. Zhang, "htSNPer1.0: software for haplotype block partition and htSNPs selection," *BMC Bioinformatics*, 6(38), 2005.

[13] V. Bafna, B. V. Halldorsson, R. Schwartz, A. G. Clark, and S. Istrail, "Haplotypes and informative snp selection algorithms: Don't block out information," *Proc. of the 7th International Conference on Computational Molecular Biology*, 2003.

[14] B. V. Halldorsson, V. Bafna, R. Lippert, R. Schwatz, F. M. D. L. Vega, A. G. Clark, and S. Istrail, "Optimal haplotype block-free selection of tagging snps for genome-wide association studies," *Genome Research*, vol. 14, 2004.

[15] E. Halperin, G. Kimmel, and R. Sharmir, "Tag snp selection in genotype data for maximizing snp prediction accuracy," *Bioinformatics*, vol. 21, no. Suppl. 1, 2005.

[16] Z. Lin and R. B. Altman, "Finding haplotype tagging snps by use of principal components analysis," *Am. J. Hum. Genet*, vol. 75, 2004.

[17] 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," *Annals of Human Genetics*, 65:95, 2001.

[18] V. North, D. Curtis, P. G. Cassell, G. A. Hitman and P. C. SHAM, "Assessing optimal neural network architecture for identifying disease-associated multi-marker genotypes using a permutation test, and application to calpain 10 polymorphisms associated with diabetes," *Annals of Human Genetics*, 67(4):348-56, 2003.

[19] A. Serretti and E. Smeraldi, "Neural network analysis in pharmacogenetics of mood disorders", *BMC Medical Genetics*, 5:27, 2004.

[20] K. C. Cartier and D. Baechle, "An artificial neural network for estimating haplotype frequencies", *BMC Genetics*, 6(Suppl 1):S129, 2005.

[21] M. Riedmiller and H. Braun, "A direct adaptive method for faster backpropagation learning: the RPROP algorithm," *Proc. of the IEEE International Conference on Neural Networks*, 1:586-591, 1993.

[22] T. M. Mitchell, *Machine Learning*, McGraw Hill, 1997.

[23] M. R. nad, S.L. Taylor, A. Clark, and D. Nicerson, "Sequence variance in the human angiotensin converting enzyme," *Nature Genetics*, vol. 22, 1999.

[24] D. Nickerson, S. Taylor, S. Fullerton, K. Weiss, A. Clark, J. Stengaard, V. Salomaa, E. Boerwinkle, and C. Sing, "Sequence diversity and large-scale typing of snps in the human apolipoprotine e gene," *Genome Research*, vol. 10, 2000.

[25] M. Daly, "High-resolution haplotype strucutre in the human genome," *Nature Genetics*, vol. 29, no. 2, 2001.

[26] M. Nei, *Molecular evolutionary genetics*, New York: Columbia University Press, 1987.

[27] M. Stephens, N. J. Smith, and P. Donnelly, "A new statistical method for haplotype reconstruction from population data", *American Journal of Human Genetics*, 68(4):978-989, 2001.

[28] G. Kimmel and R. Shamir, "GERBIL: Genotype resolution and block identification using likelihood", *Proc. of the National Academy of Sciences of the United States of America*, 102(1):158-162, 2005.

[29] D. B. Goldstein, K. R. Ahmadi, M. E. Weale, and N. W. Wood, "Genome scans and candidate gene approaches in the study of common diseases and variable drug responses," *Trends in Genetics*, 19(11):615-622, 2003.

[30] X. Ke, C. Durrant, A. P. Morris, S. Hunt, D. R. Bentley, P. Deloukas and L. R. Cardon, "Efficiency and consistency of haplotype tagging of dense SNP maps in multiple samples," *Human Mol. Gen.*, 21(13):2557-2565, 2004.

[31] P. A. Doris, "Hypertension genetics, single nucleotide polymorphisms, and the common disease: common variant hypothesis," *Hypertension*, 39:323-331,2002.

[32] L. Kruglyak and D. A. Nickerson, "Variation is the spice of life," *Nature Genetics*, 27:234-236, 2001.