

# Haplotype Tagging using Support Vector Machines

Jingwu He, Jun Zhang, Gulsah Altun, Alexander Zelikovsky and Yanqing Zhang

Department of Computer Science

Georgia State University, Atlanta, Georgia 30303

{jingwu, jun, gulsah, alexz and yzhang}@cs.gsu.edu

**Abstract**— Constructing a complete human haplotype map can help in associating complex diseases with SNPs (single nucleotide polymorphisms). Unfortunately, the number of SNPs is very large and it is costly to sequence many individuals. Therefore, it is desirable to reduce the number of SNPs that should be sequenced to a small number of informative representatives called *tag SNPs*. Depending on the application, tagging can achieve either budget savings by inferring non-tag SNPs from tag SNPs or shortening lengthy and difficult to handle SNP sequences obtained from Affimetrix Map Array. Tagging should first choose which SNPs to use as tags and then predict the unknown non-tag SNPs from the known tags.

In this paper we propose a new SNP prediction using a robust tool for classification – Support Vector Machine (SVM). For tag selection we use a fast stepwise tag selection algorithm. An extensive experimental study on various datasets including 3 regions from HapMap shows that the tag selection based on SVM SNP prediction can reach the same prediction accuracy as the methods of Halldorson et al. [7] on the LPL using significantly fewer tags. For example, our method reaches 90% SNP prediction accuracy using only 3 tags for Daly et al. [6] dataset with 103 SNPs. The proposed tagging method is also more accurate (but considerably slower) than multivariate linear regression method of He et al. [12].

## I. INTRODUCTION

In diploid organisms each chromosome has two “copies” which are not completely identical. Each of two single copies is called a haplotype, while a description of the data consisting of mixture of the two haplotypes is called a genotype. For complex diseases caused by more than a single gene it is important to obtain haplotype data which identify a set of gene alleles inherited together. In haplotype description it is important only positions where the two copies are different which are called single nucleotide polymorphisms (SNPs). A SNP is a single nucleotide site where exactly two (of four) different nucleotides occur in a large percentage of the population. Usually, a genotype is represented by a vector with coordinates 0,1, or 2, where 0 represents the homozygous site with major allele, 1 represents the homozygous site with minor allele, and 2 represents the heterozygous site. Respectively, each haplotype’s coordinate is 0 or 1, where 0 represents the major allele and 1 represents the minor allele.

Genome-wide SNP scans for disease association tests are still infeasible. In order to decrease SNP genotyping cost it is quite attractive to sequence only small amount of SNPs, so called tag SNPs, and then infer the rest of SNPs (or certain suspicious SNP’s) based on the sequenced tag SNPs. Since the SNPs responsible for complex diseases are unknown, the tag SNPs should allow to reconstruct all (or almost all) SNPs.

Note that complete 100% correct reconstruction is impossible just because a single mutation may spoil otherwise reliable reconstruction.

This problem can be formulated as **Haplotype Tagging Problem** (see Figure 1). Given the full pattern of all haplotypes in a small population sample, find the minimum number of tag SNPs and the method for reconstructing each haplotype in the entire population from these tags.

The corresponding **SNP prediction problem** is formulated as follows: Given the values of  $k$  tags of the individual  $x$  with unknown SNP  $s$  and  $n$  individuals with  $k$  tag SNP and known value of SNP  $s$ , find the value of  $s$  in  $x$  (See Figure 2).

In this paper we propose a new SNP prediction using a robust tool for classification – Support Vector Machine (SVM). For tag selection we use a fast stepwise tag selection algorithm. An extensive experimental study on various datasets including 3 regions from HapMap shows that the tag selection based on SVM SNP prediction can reach the same prediction accuracy as the methods of Halldorson et al. [7] on the LPL using significantly fewer tags. For example, our method reaches 90% SNP prediction accuracy using only 3 tags for Daly et al. [6] dataset with 103 SNPs. The proposed tagging method is also more accurate (but considerably slower) than multivariate linear regression method of He et al. [12].

The rest of the paper is organized as follows. Section II describes an universal methods for tag selection based on known prediction method, i.e., a fast stepwise tag selection algorithm. Section III proposes a new SNP prediction algorithm using SVM. Section IV presents the results of empirical study of suggested method. Finally, conclusion is given in Section V.

## II. STEPWISE TAG SELECTION

In this section we show how to separate the tag selection from SNP prediction, suggest a stepwise tag selection based on prediction.

Assuming self-similarity of data, one can expect that an algorithm predicting with high accuracy SNPs of an unknown individual will also predict with high accuracy SNPs of the sampled individual. Then, we expect that the better prediction algorithm will have fewer errors when predicting SNPs in the sample  $S$ . This expectation allows us to find tags using prediction algorithm as follows: We can check each  $k$ -tuple of tags and choose the  $k$ -tuple with the minimal number of errors in predicting the non-tag SNPs in the sampled individuals. Even though the sample elements are completely typed, prediction algorithms can make still errors because the

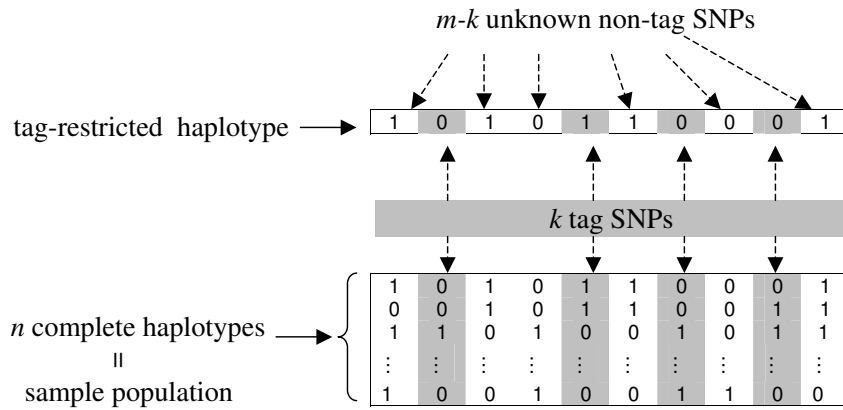


Fig. 1. Haplotype Tagging Problem. The shaded columns correspond to  $k$  tag SNPs and the clear columns correspond to non-tag SNPs. The unknown  $m - k$  non-tag SNP values in tag-restricted haplotype (top) are predicted based on the known  $k$  tag values and complete sample population.

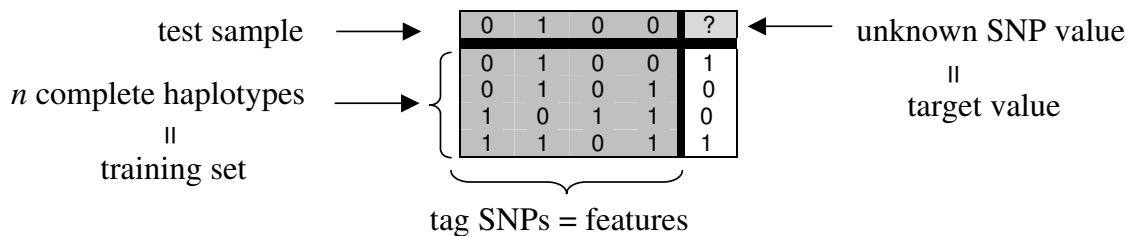


Fig. 2. The SNP Prediction Problem. Each haplotype with  $k$  tags in the training set belongs to a binary class, 0 or 1. These binary class values are represented in the last SNP column. Given a test sample of  $k$  tag-restricted haplotype, the unknown non-tag SNP in the right corner should be classified based on the known tag SNP values and training set.

number of SNPs may be not sufficient to distinguish any two sampled individuals. Thus, tag SNP selection based on prediction is reduced to the following problem:

**Tag SNP Selection.** Given a prediction algorithm  $A_k$  and a sample  $S$ , find  $k$  tags such that the prediction error  $e$  of  $A_k$  averaged over all SNPs in  $S$  (including tags) is minimized.

We propose to apply the following *Stepwise Tagging algorithm* (STA). STA starts with the best tag  $t_0$ , i.e., the SNP minimizing the error when predicting alone all other SNPs. Then STA finds such tag  $t_1$ , which would be the best extension of  $\{t_0\}$ , and continues adding best tags until reaching the set of tags of the given size  $k$ . The runtime of STA is  $O(knmT)$ , where  $T$  is the runtime of a SNP prediction algorithm. Note that STA produces a *hereditary* set of tags, i.e., the chosen  $k$  tags contain the chosen  $k - 1$  tags. The hereditary property of chosen tags allows to extend without retyping the set of tags in case of obtaining additional funding.

In this paper, we combine Support Vector Machine SNP Prediction (SVM) mentioned in Section III with Stepwise Tagging Algorithm (STA), namely SVM/STA, for solving haplotype tagging problem. We will discuss the experimental results in next section.

### III. SNP PREDICTION USING SUPPORT VECTOR MACHINE

Given the values of  $k$  tags of an unknown individual  $x$  and the known full sample  $S$ , a SNP prediction algorithm  $A_k$

predicts the value of a single non-tag SNP  $s$  in  $x$  (if there is more than one non-tag SNP to predict, then we handle each one separately). Therefore, without loss of generality, we assume each individual has exactly  $k + 1$  SNPs. Thus,  $A_k$  can be viewed as an algorithm for binary classification of vectors with  $k$  coordinates.

In this paper we propose to use Support Vector Machine (SVM) for SNP prediction. SVM has recently attracted a lot of attention in bioinformatics research. This is because SVM produces very accurate results and highly competitive with other data mining approaches such as Neural Networks. The SVM method is a learning system which is developed by Vapnik and Cortes [16]. SVM is a powerful methodology for solving problems in nonlinear classification, function estimation and density estimation. The basic principle behind SVM is to find an optimal maximal margin separating hyperplane between two classes. The goal is to maximize the margin between the solid planes separating the two classes and at the same time permit the least amount of errors as possible. SVM can also be used in the case when the data is not linearly separable. In this case, the data is mapped to a high dimensional feature space using a nonlinear function. When using SVM, the dot products  $(x,y)$  in the feature space must be fed to the SVM, which can be computed through a positive definite kernel in the input space.

After given a training set (a set of pairs, input vector:

TABLE I

LEAVE-ONE-OUT TESTS ARE PERFORMED ON 3 REAL HAPLOTYPE DATASETS. THE MINIMUM NUMBER OF TAG SNPs NEEDED TO REACH FROM 80% TO 99% PREDICTION ACCURACY IS LISTED. THE BOLT NUMBERS INDICATE CASES WHEN THE SVM/STA NEEDS FEWER TAGS THAN THE MLR METHOD OF HE ET AL. [12] FOR REACHING SAME PREDICTION ACCURACY.

datasets (num of SNPs)	prediction accuracy %											
	80	85	90	91	92	93	94	95	96	97	98	99
5q31 (103)	1	<b>1</b>	<b>3</b>	<b>3</b>	<b>4</b>	<b>5</b>	<b>6</b>	<b>8</b>	<b>10</b>	<b>22</b>	<b>42</b>	<b>51</b>
TRPM8 (101)	1	<b>1</b>	<b>2</b>	5	5	6	7	8	10	15	15	24
STEAP (22)	1	1	1	1	1	1	1	2	<b>2</b>	<b>2</b>	<b>2</b>	2

TABLE II

THE COMPARISON OF OUR PROPOSED SVM/STA METHOD AND THE MLR METHOD OF HE ET AL. [12] OVER DIFFERENT NUMBER OF TAG SNPs.

datasets (num of SNPs)	methods	number of tag SNPs						
		1	2	4	6	8	10	
5q31 (103)	prediction accuracy %	SVM/STA	86.81	89.32	92.24	94.09	95.28	96.09
		MLR	81.15	83.84	88.15	90.91	92.66	93.49
	running time	SVM/STA	3 hour	5 hour	11 hour	16 hour	18 hour	1 day
		MLR	0.77 sec	1.16 sec	4.07 sec	7.27 sec	11.26 sec	15.92 sec
TRPM8 (101)	prediction accuracy %	SVM/STA	88.89	90.50	90.67	93.67	95.56	96.74
		MLR	80.68	85.32	90.75	93.74	95.16	96.38
	running time	SVM/STA	1 hour	2 hour	5 hour	9 hour	16 hour	23 hour
		MLR	0.357 sec	0.787 sec	1.895 sec	3.376 sec	5.181 sec	7.373 sec
STEAP (22)	prediction accuracy %	SVM/STA	94.02	98.18	99.68	99.73	99.79	99.80
		MLR	90.79	96.16	99.13	99.71	99.78	99.78
	running time	SVM/STA	14 min	27 min	1 hour	2 hour	3 hour	4 hour
		MLR	0.034 sec	0.052 sec	0.118 sec	0.203 sec	0.304 sec	0.413 sec

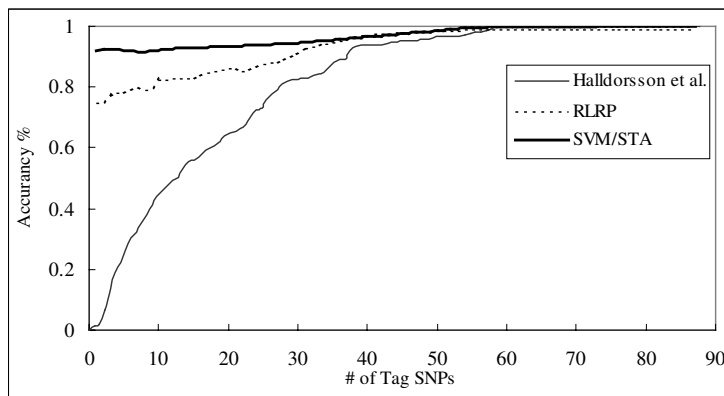


Fig. 3. Comparison among three haplotype tagging method on LPL data: SVM/STA, Halldorsson et al. [7], and He et al. [12] in a leave-one-out experiment. The x-axis shows the number of SNPs typed, and the y-axis shows the fraction of SNPs correctly imputed.

features and target), SVM builds a model. This model is later applied to unknown test set where the model maps an input vector to +1 (positive class) or -1 (negative class) output target value. In the SNP Prediction Problem, SVM builds a model after given n complete haplotypes as training set. Then when an unknown haplotype is given to SVM as a test sample, SVM is asked to predict the unknown SNP value (see Figure 2).

SVMLight is an implementation of Vapnik's Support Vector Machine [15]. In this project, we have used *SVMLight* software as a black box to do the prediction. The *SVMLight* software has many features such as changing the kernel function and other parameters. We have used the Radial Basis Function (RBF) kernel in our project it is the default and

recommended kernel function.

$$\exp(-\gamma * |u - v|^2)$$

For the trade-off between training error and margin, 0.05 is chosen (c value). Parameter gamma in RBF kernel was chosen as 0.1. These parameters were found by using try and error in our experiments and once the optimal parameters were found, we used the same for all the tests.

#### IV. EXPERIMENTAL RESULTS

We apply our haplotype tagging algorithm (SVM/STA) to very well known haplotype datasets. These datasets are original genotype datasets, but we phased them to obtain haplotypes using GERBIL algorithms [13].

**Two gene Regions form HapMap.** Two gene regions STEAP and TRPM8 from 30 CEPH family trios are obtained from HapMap [1]. We took the HapMap SNPs that are spanned by the gene plus 10KB upstream and downstream. The number of SNPs genotyped in each gene region is 23 and 102 SNPs. We only use 60 haplotypes of parents.

**Chromosome 5q31.** The data set collected by Daly et al. [6] is derived from the 616 kilobase region of human Chromosome 5q31 that may contain a genetic variant responsible for Crohn's disease by genotyping 103 SNPs for 129 trios. We only use 258 haplotypes of offsprings.

**LPL** The Clark et al. [5] data set consists of the haplotypes of 71 individuals typed over 88 SNPs in the human lipoprotein lipase (LPL) gene.

We apply leave-one-out cross-validation to evaluate the quality of the solution given by the tag SNP selection and prediction methods. One by one, each individual is removed from the sample. Then, tag SNPs are selected using only the remaining individuals. The "left out" individual is reconstructed based on its tag SNPs and the remaining individuals in the sample. The average number of errors in the reconstruction of all individuals is used as a measure of the overall prediction accuracy.

Table I presents the results of STA combined with SVM (SVM/STA) on leave-one-out experiments on the 3 haplotype datasets. Table II compares SVM/STA with multivariate linear regression method (MLR) of He et al. [12] on the 3 haplotype datasets. The proposed tagging method is more accurate than multivariate linear regression method of He et al. [12]. For example, for small number of tag SNPs, SVM/STA can obtain (up to 8%) better prediction accuracy than MLR with same number of tag SNPs. But SVM/STA is considerably slower. Indeed, for 5q31 dataset, SVM/STA needs 3 hours to select 1 tag SNPs while MLR only needs 0.77 seconds. All experiments are performed on a computer with Intel Pentium 4, 3.06Ghz processor and 2 GB of RAM.

We also compare SVM/STA with the methods of Halldorson et al. [7] and the method of He et al. [12] in leave-one-out tests on the LPL data set (see Figure 3). Note that the method of Halldorson et al. imputes a SNP based on the tag SNPs in the same neighborhood and in fact can be classified as a method for statistical coverage. If there is no tag SNPs in the neighborhood, then their method does not make any prediction. It is not surprising that it performs poorly for SNP prediction. The SVM/STA method reconstructs each SNP based on the values of *all* tag SNPs which may potentially be far away. On the LPL dataset, SVM/STA reaches, e.g., 90% accuracy using only one tag.

## V. CONCLUSIONS AND FUTURE WORK

In this paper we propose a new SNP prediction using Support Vector Machine (SVM) which is a robust tool for classification, prediction and regression in noisy, complex domains while for tag selection we use a fast stepwise tag selection algorithm. The experimental study on various datasets including 3 regions from HapMap shows that the tag selection

based on SVM SNP prediction can reach the same prediction accuracy as the methods of Halldorson et al. [7] on the LPL using significantly fewer tags. The proposed tagging method is also more accurate (but considerably slower) than multivariate linear regression method of He et al. [12]. In our future work, we will explore possibility of apply SVM/STA on genotype data by using multiclass SVM.

## ACKNOWLEDGMENT

The authors would like to thank Prof. T. Joachims for making SVMlight software available. This research was supported in part by the P20 GM065762- 01A1. Jingwu He are supported by Georgia State University Molecular Basis of Disease Fellowship.

## REFERENCES

- [1] <http://www.hapmap.org>
- [2] Avi-Itzhak, H.I., Su, X., and de la Vega, F.M. Selection of minimum subsets of single nucleotide polymorphism to capture haplotype block diversity. In Proceedings of Pacific Symposium on Biocomputing, vol. 8, pp. 466-477, 2003.
- [3] Bafna, V., Halldorsson, B.V., Schwartz, R.S., Clark, A.G., and Istrail, S. Haplotypes and informative SNP selection algorithms: don't block out information. Proceedings of the Seventh International Conference on Research in Computational Molecular Biology, pp. 19-27, 2003.
- [4] Carlson C.S., Eberle M. A., Rieder M. J., Yi Q., Kruglyak L., and Nickerson D. A. Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. American Journal of Human Genetics, vol. 74 no.1 pp.106-120, 2004.
- [5] Clark, A., Weiss, K., Nickerson, D., Taylor, S., Buchanan, A., Stengard, J., Salomaa, V., Vartiainen, E., Perola, M., Boerwinkle, E., et al. Haplotype structure and population genetic inferences from nucleotide-sequence variation in human lipoprotein lipase. Am. J. Hum. Genet. vol. 63 pp. 595-612, 1998.
- [6] Daly M., Rioux J., Schaffner S., Hudson T., and Lander E. High resolution haplotype structure in the human genome. Nature Genetics, vol 29, pp. 229-232, 2001.
- [7] Halldorsson B.V., Bafna V., Lippert R., Schwartz R., De La Vega F.M., Clark A.G., Istrail S. Optimal haplotype block-free selection of tagging SNPs for genome-wide association studies. Genome Res. vol. 14 no.8 pp.1633-40, 2004.
- [8] Halperin, E., Kimmel, G. and Shamir, R. "Tag SNP Selection in Genotype Data for Maximizing SNP Prediction Accuracy", *Bioinformatics* 21:i195-i203, 2005.
- [9] He J. and Zelikovsky A. Linear Reduction Methods for Tag SNP Selection, Proc. International Conf. of the IEEE Engineering in Medicine and Biology (EMBC'04), pp. 2840-2843, 2004.
- [10] He J. and Zelikovsky A. Linear Reduction for Haplotype Inference. Proc. Workshop on Algorithms in Bioinformatics (WABI'04), Lecture Notes in Bioinformatics (LNBI) 3240, pp. 242-253, 2004.
- [11] He, J. and Zelikovsky, A. 'Linear Reduction Method for Predictive and Informative Tag SNP Selection', *International Journal Bioinformatics Research and Applications*, vol. 3, pp. 249-260, 2005.
- [12] He, J. and Zelikovsky, A. "Tag SNP Selection Based on Multivariate Linear Regression", *International Workshop on Bioinformatics Research and Applications (IWBRA 2006)*, Manuscript.
- [13] Kimmel, G and Shamir R. "GERBIL: Genotype resolution and block identification using likelihood", *PNAS* vol. 102, pp. 158-162, 2004.
- [14] Sebastiani, P., Lazarus, R., Weiss, S., Kunkel, L., Kohane, I., and Ramoni, M. Minimal haplotype tagging. Proc. Natl. Acad. Sci. vol. 100 pp. 9900C9905, 2003.
- [15] T. Joachims, Making large-Scale SVM Learning Practical. Advances in Kernel Methods - Support Vector Learning, B. Scholkopf and C. Burges and A. Smola (ed.), MIT-Press, 1999.
- [16] V. Vapnik and C. Cortes, support vector networks, Machine Learning, vol. 20, pp. 273-293, 1995.
- [17] Zhang, K., Qin, Z., Liu, J., Chen, T., Waterman, M., and Sun, F. Haplotype block partitioning and tag SNP selection using genotype data and their applications to association studies. Genome Res. vol. 14 pp. 908C916, 2004.