NEW KERNEL METHODS FOR PHENOTYPE PREDICTION FROM GENOTYPE DATA

RITSUKO ONUKI[1] TETSUO SHIBUYA[2]
onuki@hgc.jp tshibuya@hgc.jp

MINORU KANEHISA[1,2]
kanehisa@kuicr.kyoto-u.ac.jp

[1] *Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokosho, Uji, Kyoto 611-0011, Japan*
[2] *Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan*

Phenotype prediction from genotype data is one of the most important issues in computational genetics. In this work, we propose a new kernel (i.e., an SVM: Support Vector Machine) method for phenotype prediction from genotype data. In our method, we first infer multiple suboptimal haplotype candidates from each genotype by using the HMM (Hidden Markov Model), and the kernel matrix is computed based on the predicted haplotype candidates and their emission probabilities from the HMM. We validated the performance of our method through experiments on several datasets: One is an artificially constructed dataset via a program GeneArtisan, others are a real dataset of the NAT2 gene from the international HapMap project, and a real dataset of genotypes of diseased individuals. The experiments show that our method is superior to ordinary naive kernel methods (i.e., not based on haplotype prediction), especially in cases of strong LD (linkage disequilibrium) .

Keywords: genotypes; SVM; phasing; single-nucleotide polymorphism.

1. Introduction

Most variations of the DNA sequences in humans are at single-base sites, in which more than one nucleic acid can be observed across the population [5]. These sites (i.e. loci) are called SNPs (single nucleotide polymorphisms). The sequence of pairs of DNA residues at each SNP site is called genotypes. The DNA residues which consist of genotypes are called alleles. There are two kinds of allele, major allele and minor allele. The major allele is an allele which is observed more frequently than another allele at the SNP site. The minor allele is less frequently observed than major allele at the SNP site. A sequence of contiguous alleles of SNPs on the same chromosome is called a haplotype. Most experimental techniques for determining SNPs do not provide haplotype information [4], [8]. The experiments generate only an unordered pair of allele readings for each site on the two chromosomes, i.e., a genotype. To obtain the haplotype data, we need to infer the haplotypes from genotype data. The process of inferring haplotypes from genotypes is called phasing or resolving.

For the tailor-made medicine, the variations of the individuals' phenotypes caused by the differences of DNA sequences become important and it is becoming important to predict individuals' phenotypes from genotypes. In this work, we developed new kernel methods to predict phenotype from genotype data, utilizing phased haplotype information. Then, we applied our methods to the disease status prediction and the prediction of the individual's geographic background. Previous work for SVM-based disease status prediction is based on the number of the minor allele for each SNP site [2]. We focused on the phased haplotype and discussed whether the phased haplotype can improve the prediction accuracies.

There are many software tools for haplotype inference: PHASE [20], [21], GERBIL [13], HAP [9] and HIT [19]. Among the tools listed above, HIT can provide not only the optimal haplotype but multiple candidates of haplotypes. Based on the multiple candidates of haplotypes obtained by HIT, we propose new kernel methods.

In the last part of the paper, we demonstrate the effectiveness of these methods through experiments on several datasets, i.e., two kinds of real genotype datasets and an artificial dataset.

2. Methods

2.1. *Preliminaries*

In this section, we describe the notations and 2 methods, i.e., Haplotype Inference Technique (HIT) [19] and haplotype similarity measure, which are the basis of our methods.

Notations

Consider n genotypes over m SNP loci from the same chromosome. These loci are numbered $1,\cdots,\mathrm{m}$ from left to right in the physical order. In most cases, only two alternative bases (i.e. alleles) occur at a SNP site. The allele for the SNP locus is an element of the set $\mathrm{A}=\{1,0\}$ where 1 and 0 refer to the most frequent allele and others at each SNP locus respectively. Sometimes the data contains missing values, and we represent them by '?'. Then a haplotype is a sequence in A^{m} (e.g. 1010111100). A genotype is a sequence of unphased (i.e., unordered) allele pairs and is defined as a sequence in A'^{m}, where $\mathrm{A}'=\mathrm{A}\times\mathrm{A}$ (e.g. $00\ 00\ 00\ 11\ 01\ 11\ 0?\ 00\ 00\ 11$). Thus a genotype data consists of $2\mathrm{n}\times\mathrm{m}$ values. Given a set of genotypes, the phasing problem is to find their corresponding most probable haplotype pairs (i.e. diplotypes) that could have generated the genotypes.

Haplotype Inference Technique (HIT)

HIT is an algorithm that uses HMM (Hidden Markov Model) [17], [18] to estimate haplotypes from genotypes. It shows relatively high accuracy among other software programs [19]. From now on, we describe how HIT infers haplotypes from genotypes. In HIT, we assume that all sites are bi-allelic.

The model parameters, i.e., transition probabilities and emission probabilities, are estimated by the EM algorithm, using only the input genotype dataset. This means the

HIT does not need any training dataset. The number of founders K is a parameter we need to set before the HMM is trained. In Rastas et al. [19], they say that the HIT gives good results if K is set to any value larger than 4. They themselves use the setting $K = 7$ in their experiments against the Daly et al.'s data [6]. So we also set K to 7 in our experiments in section 3. Once the model has been trained, we can estimate haplotypes from genotypes. Moreover we can obtain multiple haplotype candidates with emission probabilities that the HMM emits the haplotype.

Haplotype similarity measure

We use the hamming distance as a similarity measure between haplotypes. For a haplotype $h \in A^m$, let $h(k)$ denote the allele at locus k of the haplotype h. The similarity measure between two haplotypes h and h' is defined as:

$$s(h,h')= \sum_{k=1}^{m} I\left(h(k),h'(k)\right) \tag{1}$$

where $I(a,b) = 0$ if alleles a and b are the same and $I(a,b) = 1$ otherwise. Tzeng et al. [22] and Li and Jiang [14] also use this hamming distance or its variant as the similarity measure between haplotypes.

To obtain length-independent measure, we consider the following value $d(h,h')$ as the similarity between haplotypes h and h':

$$d(h,h') = \frac{s(h,h')}{m}. \tag{2}$$

This measure may not be always best for all evolutionary or practical scenarios. It should be noted that our methods described in section 2.2 can use any other similarity measures between haplotypes, though there are not known any standard measures other than the measure described above.

2.2. *Our method*

In this section, we first introduce genotype-genotype distance defined in our another work [15] for computing kernels between genotypes (Section 2.2.1). Based on the kernels, we next describe how we predict phenotypes from genotype data in Section 2.2.2.

2.2.1. *Genotype distance*

To develop new kernels, we need somewhat distance between genotypes. We introduce genotype-genotype distance, haplotype frequency-based distance (HFD), proposed in [15].

Before defining the genotype-genotype distance, we define a distance between haplotype pairs based on the haplotype distance described in Section 2.1. Let $a = (h_1, h_2)$ and $a' = (h_1', h_2')$ be two haplotype pairs to be compared, where $h_1, h_2, h'_1, h'_2 \in A^m$. We define the distance between haplotype pairs a and a' as:

$$\mathrm{H}(\mathrm{a},\mathrm{a}') = \min\{\, \mathrm{d}(\mathrm{h}_1,\mathrm{h}_1') + \mathrm{d}(\mathrm{h}_2,\mathrm{h}_2'), \qquad \mathrm{d}(\mathrm{h}_1,\mathrm{h}_2') + \mathrm{d}(\mathrm{h}_2,\mathrm{h}_1')\}. \tag{3}$$

Using this distance, we propose the following genotype-genotype distance.

Haplotype frequency-based distance (HFD)

For genotypes $\mathrm{g},\mathrm{g}' \in \mathrm{A}'^{\mathrm{m}}$, let $\mathrm{c}_\mathrm{i} = (\mathrm{h}_{\mathrm{i}1},\mathrm{h}_{\mathrm{i}2})$ $(1 \le \mathrm{i} \le M)$ and $\mathrm{c}_\mathrm{j}' = (\mathrm{h}_{\mathrm{j}1}',\mathrm{h}_{\mathrm{j}2}')$ $(1 \le \mathrm{j} \le M')$ be candidate haplotype pairs for g and g' respectively, which are computed by HIT. Note that these candidate sets are not all the set of possible candidates, as there are usually an exponential number of candidates when we infer haplotypes from genotypes. The M and M' are user-specified upperbounds of the numbers of candidates to be enumerated by the HIT. In the experiments in Section 4, we let $\mathrm{M} = \mathrm{M}' = 100$.

Let p_i, p_j' be the emission probabilities of the candidate haplotype pairs, c_i and c_j'. The emission probabilities are considered as the haplotype frequencies. The haplotype frequency-based distance (HFD) between genotypes is defined as the following summation:

$$\mathrm{HFD}(\mathrm{g},\mathrm{g}') = \sum_{\mathrm{j}=1}^{M'} \sum_{\mathrm{i}=1}^{M} \mathrm{H}\big(\mathrm{c}_\mathrm{i},\mathrm{c}_\mathrm{j}'\big) \cdot \mathrm{q}_\mathrm{i} \cdot \mathrm{q}_\mathrm{j}' \tag{4}$$

where $\mathrm{q}_\mathrm{i} = \mathrm{p}_\mathrm{i}/\sum_{\mathrm{k}=1}^{M} \mathrm{p}_\mathrm{k}$ and $\mathrm{q}_\mathrm{j}' = \mathrm{p}_\mathrm{j}'/\sum_{\mathrm{k}=1}^{M'} \mathrm{p}_\mathrm{k}'$. q_i and q_j' are the normalized emission probabilities of the candidate haplotype pairs, c_i and c_j', respectively.

2.2.2. *Our learning algorithm*

We propose a learning algorithm for phenotype prediction from genotypes which consists of 3 steps. The learning algorithm needs a training dataset and a test dataset. We train our learning algorithm with the training dataset and test the leaning algorithm with the test dataset. The details of each step are described as follows.

Step1. Haplotype Inference

For each of the given training set and test set of genotypes, we extract candidates of haplotype pairs and their corresponding emission probabilities by using HIT. In the experiments in section 3, we let the number of the founders for the HMM be 7, which is the same as the setting in the experiments in Rastas et al. [19].

Step2. Computation of the kernels

Using the result of the step 1, we evaluate HFD against all the pairs of genotypes of the training set and the test set. The details of HFD are described in the previous section 2.2.1. Based on the HFD, we compute the kernels in the 2 ways. One is $\mathrm{e}^{-\mathrm{HFD}\,(\mathrm{g},\mathrm{g}')}$ for genotypes g and g', which we call exponential HFD and another is $1 - \mathrm{HFD}(\mathrm{g},\mathrm{g}')$ for genotypes g and g', which we call linear HFD. In most cases, our kernels are positive-semi definite. If the kernels are not the positive-semi definite, we do as the follows.

Let K be the kernel matrix. K can be written as $K = PMP^{-1}$, where M is a triangular matrix and its diagonal elements are the eigenvalues of K, and P is an orthogonal matrix. We replace the negative diagonal elements of M by zero and let it be M'. Let $K' = PM'P^{-1}$. We use this K' as the kernel matrix.

Step3. Phenotype Prediction by SVM

We predict the phenotypes from the test set of genotype data with SVM [24] based on the kernels computed in the step 2 of our learning algorithm.

3. Data sets

3.1. *The NAT2 dataset*

In the next section, we do experiments against a set of NAT2 gene-related genotypes [3] taken from the HapMap datasets of the version on March 1, 2007 [11]. The data consists of 270 genotypes with 24 SNP sites. The 270 genotypes can be divided into 3 ethnic groups (i.e. populations). The first group consists of genotypes of 90 Utah residents with ancestry from northern and western Europe (CEU), the second group consists of genotypes of 45 unrelated individuals of Han Chinese in Beijing and 45 unrelated individuals of Tokyo, Japan (CHB+JPT), and the last group consists of genotypes of 90 Yoruba people in Ibadan Nigeria, West Africa (YRI).

The NAT2 gene are said to be related to the susceptibility to some toxicities and cancers [7], [10]. Thus it is very important to study the differences of the NAT 2 genes among different populations to elucidate the ethnic difference in such susceptibilities.

3.2. *Disease datasets*

Dataset of Crohn's disease [6] is derived from the 616 kb region of human Chromosome 5q31 that may contain a genetic variant responsible for Crohn's disease by genotyping 103 SNPs for 129 trios. All offspring belong to the case population, while almost all parents belong to the control population. In entire data, there are 144 case and 243 control individuals.

Dataset of autoimmune disorder [23] is sequenced from 330kb of human DNA containing gene CD28, CTLA4 and ICOS. These genes are proved related to autoimmune disorder. A total of 108 SNPs were genotyped in 384 cases of autoimmune disorder and 652 controls.

3.3. *Simulated dataset*

We also applied our algorithms to an artificial genotype dataset to examine the performance of our methods. The data is generated by GeneArtisan [25] based on the

Coalescent model. The region includes one SNP site related to a genetic risk factor for diseases.

We made training datasets of 1000 genotypes, where 500 are affected (i.e. case) genotypes and 500 are normal (i.e. control) genotypes, using the above tool.

4. Results and Discussions

In the following sections, we show the results of the 3-fold cross-validations. Throughout the prediction validations, we compare our methods with other methods, i.e. Brinza' method [2], SVM-fisher [12] and the method based on the number of the major allele. SVM-Fisher is a HMM based method. We used the same HMM in the step 1 of our algorithm for SVM-Fisher. The method based on the number of the major allele counts the number of the major allele at each SNP site and makes the feature vector of each individual's genotype.

4.1. *Results on high LD datasets*

Our HFD based methods show high accuracy on YRI and CEU datasets from HapMap datasets (Table 1, Table 2). We found that the datasets which our methods show high accuracies have high Linkage Disequilibrium (LD) (Figure 1). The LD measure a co-segregation between the SNP sites. It is estimated that the haplotypes of high LD datasets can be inferred more precisely than low LD datasets (i.e. CHB+JPT dataset) and our HFD based methods show high accuracies than the other methods. Our methods also show high accuracy on the simulated datasets. The simulated datasets were generated by the tool incorporating that the human genome consists of haplotype blocks [16]. The regions in the haplotype blocks show high LD. It is estimated that the simulated datasets tend to show the haplotype structure more clearly and the haplotypes can be inferred precisely than real datasets and accuracy of our HFD based method are the higher than any other methods. And the accuracy is the higher value than the other real datasets (Table 3).

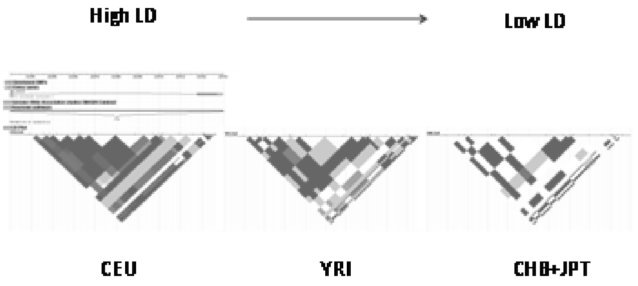

Figure 1. LD plots of NAT2 gene from HapMap by Haploview [1]. The black color and gray color show the high LD score, and the white color shows the lower LD score.

Table 1. Results of 3-fold cross-validation for CEU dataset.

Method	Sensitivity	Specificity	Accuracy
Exp HFD	0.822	0.828	0.832
Linear HFD	0.822	0.811	0.815
Major allele	0.811	0.800	0.804
SVM-Fisher	0.678	0.728	0.711
Brinza’s method	0.733	0.755	0.748

Table 2. Results of 3-fold cross-validation for YRI dataset .

Method	Sensitivity	Specificity	Accuracy
Exp HFD	0.722	0.939	0.867
Linear HFD	0.722	0.939	0.867
Major allele	0.733	0.911	0.863
SVM-Fisher	0.489	0.861	0.737
Brinza’s method	0.667	0.856	0.793

Table 3. Results of 3-fold cross-validation for simulated dataset.

Method	Sensitivity	Specificity	Accuracy
Linear HFD	0.960	1.000	0.980
Major allele	0.970	0.920	0.945
SVM-Fisher	0.980	0.620	0.800

4.2. *Results on low LD datasets*

As described in section 4.1, CHB+JPT dataset from HapMap shows lower LD than the other HapMap datasets, which we estimated is the reason why our HFD based methods don’t show higher accuracies than the other methods (Table 4). Our HFD based methods also don’t show high accuracies on the disease datasets (Table 5, Table 6). The gene related to the disease datasets are ranged in longer regions. The datasets are truncated by three haplotype blocks and the regions contain much lower LD [23], which we estimated is the reason why our HFD based methods don’t show high accuracies.

Table 4. Results of 3-fold cross-validation for CHB+JPT dataset.

Method	Sensitivity	Specificity	Accuracy
Exp HFD	0.911	0.800	0.832
Linear HFD	0.911	0.783	0.826
Major allele	0.922	0.744	0.804
SVM-Fisher	0.767	0.728	0.778
Brinza’s method	0.911	0.800	0.837

Table 5. Results of 3-fold cross-validation for Crohn’ s disease dataset.

Method	Sensitivity	Specificity	Accuracy
Exp HFD	0.468	0.603	0.543
Linear HFD	0.561	0.536	0.545
Major allele	0.574	0.456	0.496
SVM-Fisher	0.668	0.352	0.447
Brinza’s method	0.388	0.575	0.501

Table 6. Results of 3-fold cross-validation for dataset of Autoimmune disorder.

Method	Sensitivity	Specificity	Accuracy
Exp HFD	0.530	0.521	0.525
Linear HFD	0.451	0.692	0.603
Major allele	0.425	0.684	0.587
SVM-Fisher	0.493	0.524	0.515
Brinza’s method	0.601	0.570	0.580

5. Results and Discussions

We proposed kernel methods based on the support vector machine. We succeeded in accurate phenotype prediction especially for the high LD datasets, i.e. CEU and YRI dataset of NAT2 gene and simulated dataset. On the other hand, our results were not so good for the low LD datasets, i.e. JPT+CHB dataset of NAT2 gene and disease datasets. As our kernels are based on phased haplotypes, we concluded that our prediction accuracies depend on the degree of the LD of the datasets. We can infer the haplotype more precisely for the high LD datasets than low LD datasets in the step 1 in our methods. The prediction accuracy of haplotype inference influences the prediction accuracy of the SVM.

In this work, we only treated the disease status prediction. We will extend our models to take into account whether the individuals have the potential of the disease or not. We will also try to take into account the variations of the haplotype blocks to consider more biological aspects into our models.

Acknowledgments

We thank Dr. Ryo Yamada for comments on various problems in genetics related to this research. We also thank Dr. Yoshihiro Yamanishi and Dr. Hisashi Kashima for very helpful discussions.

References

[1] Barrett, J.C., Fry, B., Maller, J., Daly, M.J., Haploview: analysis and visualization of LD and haplotype maps, *Bioinformatics*, 21:263-265, 2004.

[2] Brinza, D., Zelikovsky, A., Discrete Methods for Association Search and Status Prediction in Genotyep Case-Control Studies, *Proc. of IEEE 7-th International symposium on Bioinformatics and Bioengineering* 270-277, 2007

[3] Butcher, N.J., Boulouvala, S., Sim, E. and Minchin, R.F., Pharmacogenetics of the arylamine N-acetyltransferases, *Pharmacogenomics*, 2:30-42, 2002.

[4] Clark, V.J., Metheny, N., Dean, M., Peterson, R.J., Statistical Estimation and pedigree analysis of CCR2-CCR5 haplotypes, *Hum Genet*, 108:484-496, 2001.

[5] Collins, F.S., et al., A DNA polymorphism discovery resource for reserch on human genetic variation, *Genome Res*., 8:1229-1231, 1998.

[6] Daly, M., Rioux, J., Hudson, T., Lander, E., High-resolution haplotype structure in human genome, *Nat Genet*, 29:229-232, 2001.

[7] Evans, D.A., N-acetyltransferase, *Pharmacol Ther* 42:157-234, 1989.

[8] Fallin, D., Schork, N.J., Accuracy of haplotype frequency estimation for biallelic loci, via expectation-maximization algorithm for unphased diploid genotype data, *Am J Hum Genet*, 67:947-959, 2000.

[9] Halperin, E., Eskin, E., Haplotype reconstruction from genotype data using imperfect phylogeny, *Bioinformatics*, 20: 104-113, 2004.

[10] Ito, T., Inoue, E., Kamatani, N., Association Test Algorithm Between a Qualitative Phenotype and a Haplotype or Haplotype Set Using Simultaneous Estimation of Haplotype Frequencies, Diplotype Configurations and Diplotype-Based Penetrances, *Genetics*, 168: 2339-2348, 2004.

[11] International HapMap Consortium, The International Hapmap Project, *Nature*, 426:789-796, 2003. http//www.hapmap.org

[12] Jakkola, T., Diekhans, M., Haussler, D., A Discriminative Framework for Detecting Remote Protein Homologies, *Journal of computational biology*, 7:95-114, 2000.

[13] Kimmel, G., Shamir, R., GERBIL: Genotype resolution and block identification using likelihood, *Proc Nat Acad Sci*, 102;158-162, 1998.

[14] Li, J., Jiang, T., Haplotype-based linkage disequilibrium mapping via direct data mining, *Bioinformatics*, 21(24):4384-4393, 2005.

[15] Onuki R., Shibuya, T., Yamada, R., Kanehisa, M., A new measure of inter-diplotype distance using haplotype frequency in populations, in preparation.

[16] Phillip, M.S., Lawrence, R., Sachidanandam, R., Morris, A.P., Balding, D.J., Donaldson, M.A., Studebaker, J.F., et al. Chromosome-wide distribution of haplotype blocks and the role of recombination hot spots, *Nat Genet*, 33:382-387, 2003.

[17] Rabiner, L.R., Juang, B.H., An Introduction to Hidden Markov Models, *IEEE ASSP Mag*., 3(1):4-16, 1986.

[18] Rabiner, L.R., A tutorial on hidden Markov models and selected applications in speech recognition, *Proceedings of the IEEE,* 77: 257-285, 1989.

[19] Rastas, P., Koivisto, M., Mannila, H., Ukkonen, E., A hidden markov technique for haplotype reconstruction, *Lecture Notes in Bioinformatics,* 3692:140-151, 2005.

[20] Stephans, M., Smith, N., Donnelly, P., 2001. A new statistical method for haplotype reconstruction from population data, *Am J of Hum Genet,* 68:978-989, 2001.

[21] Stephans, M., Scheet, P., Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation, *Am J of Hum Genet,* 76:449-462, 2005.

[22] Tzeng, J.Y., Devlin, B., Wasserman, L., Roeder, K., On the identification of disease mutations by the analysis of haplotype similarity and goodness of fit, *Am J of Hum Genet,* 72:891-902, 2003.

[23] Ueda, H., Howson, J.M.M., Esposito, L., et al. 2003. Association of the T cell Regulatory Gene CTLA4 with Susceptibility to Autoimmune Disease, *Nature*, 423 506-511, 2003.

[24] Vapnik, V., 1998. Statistical Learning Theory, *Wiley, NY,* 1998.

[25] Wang, Y., Rannala, B., *In Silico* Analysis of Disease-Association mapping Strategies Using the Coalescent Process and Incorporating Ascertainment and Selection, *Am J of Hum Genet,* 76:1066-1073, 2005.