SNP Auto-Calling Using Support Vector Machines

Kai Zhang[1] Marc Ma[1, 2] Hui-Yun Wang[3] Yu Wang[1] Malini Banerjee[1] Avik Karmaker[1] Frank Shih[1] Jason Wang[1] Honghua Li[3]

[1]Dept. of Computer Science, [2]Center of Applied Mathematics and Statistics, New Jersey Institute of Technology
[3]Dept. of Molecular Genetics and Microbiology/The Cancer Institute of New Jersey, University of Medicine and Dentistry of New Jersey, Robert Wood Johnson Medical School
Emails: {kxz6841, qma}@njit.edu

Abstract

Technological breakthroughs have enabled researchers to amplify thousands of DNA sequences containing single nucleotide polymorphisms (SNPs) in a single multiplex polymerase chain reaction (PCR). The genotypes of SNPs can then be determined by microarray assay. The enormous amount of raw data produced in routine microarray experiments precludes manual processing to generate genotype calls. In this paper, we present a novel algorithm, GenoIterSVM, to perform automatic genotype-calling. GenoIterSVM uses Support Vector Machines (SVMs) as the algorithmic basis and introduces iterative refinement for improved classification accuracy. Thousands of SNPs have been interrogated using our PCR-based genotyping system. Auto-calling of the genotypes of these SNPs was performed using our GenoIterSVM with high accuracy and efficiency, which renders GenoIterSVM suitable for auto-calling in high-throughput genotyping microarray experiments.

Keywords: single nucleotide polymorphism (SNP), automatic genotype-calling, high-throughput genotyping microarrays, support vector machine (SVM), multiplex polymerase chain reaction (PCR).

1. Introduction

Single nucleotide polymorphism (SNP) is the most common genetic variation in human genome, at a frequency of one SNP per 600 base pairs (bp) [1]. More than 20,000,000 SNPs have been deposited in dbSNP, a SNPs database maintained by National Center for Biotechnology Information with the completion of human genome project [2]. New diagnostic and treatment methods for many complex diseases can be developed using these SNPs as genetic markers [3].

Many technologies have been developed to explore the information of SNPs, such as Affymatrix genotyping system, Ilumina BeadArray genotyping system and Molecular Inversion Probe (MIP) genotyping system. Recently, we also have developed a genotyping system, which is based on multiplex polymerase chain reaction (PCR) and microarrays [4], Using our unique genotyping system, more than one thousand of sequences containing SNPs sites can be amplified in a single tube and the genotypes can be interrogated using one microarray [4].

These high-throughput genotyping microarray systems generate intimidating amount of raw data in routine experiments, which precludes the feasibility of manual processing. Automated procedures and efficient algorithms are sorely needed to perform auto-calling to determine genotypes. In exploration and analysis of such massive microarray data, many machine learning and data mining techniques have been developed, such as the Modified Partitioning Around Medoids (MPAM) method for Affymetrix genotyping microarrays [5]. However, methods like MPAM are specifically designed for Affy data analysis, and therefore not suited to the analysis of the data generated in our genotyping system [4]. We have developed a support vector machine (SVM) based method to perform auto-calling of genotypes of SNPs. In our auto-calling method, we also introduce iterative refinement in the learning and testing stages for improved classification accuracy. The resulting algorithm is termed as GenoIterSVM.

SVM is a widely used supervised machine learning method [6]. It was first introduced to perform binary classification, *i.e.*, determining which class a given data point belongs to. Recent advances have rendered SVMs capable of performing multi-class classification. An SVM determines the parameters of a single learning unit through a virtual nonlinear projection of the input data into a feature space with higher dimension. It guarantees to find the optimal hyper-plane that is least likely to overfit, which makes SVM advantageous in many applications. The one-against-one method is an extension of SVM for multi-class classification, in which $k(k-1)/2$ SVMs are constructed in a tree structure. Each SVM represents a distinguishable pair-wise classifier from different

classes in the training stage, in which one needs to solve the following optimization problem:

$$\min(\frac{1}{2} w^{ij^T} w^{ij} + C\sum_{t=1}^{l} \xi_t^{ij}),$$

where w^{ij} is the slope of the hyper-plane separating i^{th} and j^{th} classes, C is soft margin parameter, and ξ_t^{ij} is the penalty. In the testing stage, each leaf SVM pops up one class label, which propagates to the upper level in the tree until it processes the root SVM of the tree.

Using an auto-calling procedure like ours, we are able to eliminate the need for human intervention. We do not need to set any empirical cutoff values. Auto-calling using GenoIterSVM achieves high accuracy, reliability and efficiency, and is suitable for the high-throughput microarray assays.

2. Multiplex PCR-Based Microarray Genotyping System

In our multiplex PCR-based microarray genotyping system, more than one thousand of SNPs can be simultaneously amplified in one reaction and the genotypes can be detected by minisequencing on a microarray [4]. We select biallelic SNPs from dbSNP (ftp://ftp.ncbi.nih.gov/snp/human/chr_rpts/), a SNP database maintained by NCBI. In two-color SNP genotyping microarray experiments on haploid samples, there are three possible genotypes for each SNP locus: homozygotes "C/C", "T/T" and heterozygote "C/T" on one strand from one direction. If genotyping the SNPs from another direction, the genotypes will be "G/G", "A/A" and "G/A". However, the analysis software will only give "C/C", "T/T" and "C/T" based on the different fluorescent (Cy3 and Cy5) intensities of the loci no matter which direction has been used for genotyping. Microarrays were scanned using GMS 418 Array Scanner, Genepix 4100 Scanner after washing off the dyes in the solution. Images are digitized using Genepix or ImaGene software package.

3. Auto-Calling Algorithms

Before making genotype calls, we need to perform preprocessing of the raw data, a crucial step towards accurate genotype-calling, which includes data normalization to remove channel imbalance and noise subtraction to reveal the true signals.

After preprocessing of the raw data, we will be ready to apply GenoIterSVM to determine the genotype for each data point. GenoIterSVM consists of a basic SVM procedure as its core and an iterative refinement procedure for improved classification accuracy. SVMs may employ different kernels to perform the nonlinear projection of the input data into a feature space. Here we use the simplest kernel, the linear kernel, for our classification tasks. Results on how accuracy is affected by choosing different kernels are reported in a separate paper. Our methods are implemented in Matlab.

3.1. Data preprocessing

Each channel may have bias in digitization of the fluorescence signals, *e.g.*, red signals may be systematically stronger than green signals [7]. The most pronounced systematic channel imbalance that does not contribute to differential expression between the two types of alleles (C and T or A and G) is the imbalance of the green and red dye incorporation. The following procedure performs channel normalization by correcting channel imbalance due to the fluorescent intensity. A more rigorous approach would correct the channel imbalance due to spatial factors as well, e.g., the novel stepwise normalization method [8]. For initial normalization [4], we sort the spots in descending order in terms of the ratio of intensities, *r/g*, from the two channels where *r* values are from mean or median intensities from red signal channel and *g* values are mean or median intensities from green signal channel. The leading and trailing spots are tentatively classified as homozygotes. Then we compute the ratio, $\bar{r}/\bar{g}$, in which $\bar{r}$ and $\bar{g}$ are the mean of median intensities of the leading and trailing spots: $(P_1, P_2, P_3, ...P_n)$, $(P_{m+1}, P_{m+2}, P_{m+3}, ...P_{m+n})$. Each data point is then normalized according to this ratio.

The raw data consist of the true signals and the noises due to different background, fluorescence remission and scanning efficiency. We need to eliminate the noises from the raw data to reveal the true signals. Following the normalization step, the average of intensities of green signals for the spots $(P_1, P_2, P_3, ...P_n)$ is used for green background, whereas the average of intensities of green signals for the spots $(P_{m+1}, P_{m+2}, P_{m+3}, ...P_{m+n})$, is used for red background. The "true" signal is computed by subtracting the background from the normalized values. After background subtraction, spots with combined intensity lower than a threshold (usually twice the standard deviation of the noises) are flagged "L" for "low-signal no-call". The novelty of this noise subtraction method is that the nonspecific hybridization effect, which is the most difficult to estimate, is factored into these estimates.

3.2. The basic SVM

A basic SVM method for genotyping data classification involves no iterative processes in learning and testing steps. For classification using basic SVM, the first stage is the supervised learning

using training data set, and the second stage is the blinded testing on the testing data set. During the learning step, the SVM learns from the preprocessed data, which are more easily separable since the channel imbalance is corrected and noise is subtracted. The learning process results in a system of classifiers that are used for blinded testing.

Fig. 1 shows the linear SVM classifier learned from a set of microarray data generated using our genotyping system [4] using the linear kernel, in which there are three regions separated by two straight lines that are not necessarily parallel to each other. The preprocessed data are transformed to the logarithmic scales to facilitate the supervised learning. These SVM classifiers can be used to perform subsequent blinded testing tasks on hard-to-separate data sets.

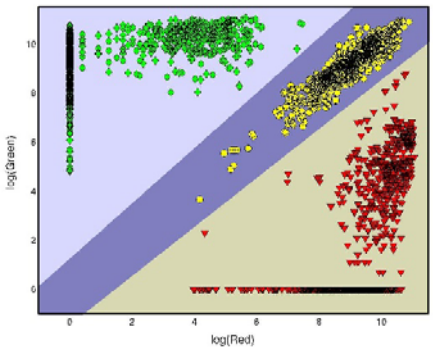

Fig. 1: SVM classifiers using the linear kernel.

Note that the separation of data into three distinct groups can be very hard or even impossible for routine genotyping microarrays even after preprocessing of raw data. Logarithmic transformation is essential to guarantee efficiency of SVM calculations, as is done in our procedures. Without logarithmic transformation the computational cost in finding the correct classifiers will be substantially higher due to the extreme large numerical values of raw data.

3.3. Iterative refinement

Classification accuracy could be improved by introducing iterative refinement in the learning and testing stages in the basic SVMs. The resulting algorithm is termed as GenoIterSVM, which consists of a basic SVM as the core and an iterative procedure for improving classification accuracy. With this algorithm, the data are iteratively re-normalized during the learning and the blinded testing stages.

In SNP genotyping microarrays, usually there are plenty data points representing heterozygous alleles at DNA loci. Ideally, these data points should distribute evenly along the bisector of the first quadrant of the coordinate system. In practice, however, this is not always true. We impose an artificial regularity constraint on the SVM classifiers: the optimal curve fit for these heterozygote data points using linear regression analysis must be collinear with the bisector of the first quadrant. If this regularity constraint is not met after the learning stage, we argue that a systematic bias is inherent in the data set and must be corrected.

Fig. 1 shows that the collinear constraint is not met for the SVM classifiers obtained after the initial learning stage – the slope of the linear curve fit for the heterozygote data points is not equal to 1, which means in general the green intensity is greater than the red intensity. We correct this systematic error iteratively by shifting and rotating the coordinate system. The iteration will be stopped when this constraint is met. This iterative learning process results in a system of canonical classifiers, as shown in Fig. 2. We use a similar iterative approach in the subsequent blinded testing step. Note that the convergence will take place after just one iteration if none of the data points used in linear curve fitting disappears from this set. It is possible that some heterozygote data points will be labeled as homozygote “C” or homozygote “T” after one such iteration. We found that in most cases it is enough to achieve convergence using one iteration.

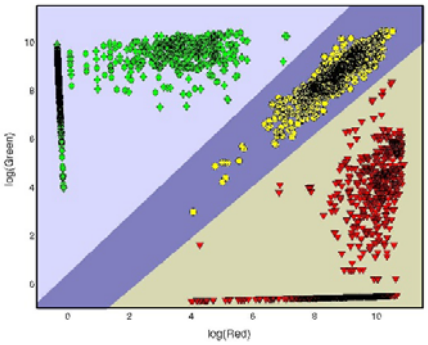

Fig. 2: SVM classifier with iterative refinement.

4. Results of auto-calling using GenoIterSVM

About 4,600 SNPs covering 12 chromosomes in at least 24 genomic DNA samples and 30 human single sperms are genotyped with high detection rate in DNA samples and single sperms, and genotypes for all experiments were determined using a simple linear cutoff-based method [4]. We also performed genotype-calling on a large subset of these raw microarray experiments using GenoIterSVM.

Though in principal we could determine the genotypes for all DNA loci using other independent genotyping methods such as Restriction Fragment Length Polymorphism (RFLP) and direct sequencing,

which can then be used to validate the classification accuracy of our auto-calling algorithms, in reality it is not practical to do so since RFLP and direct sequencing methods are very time-consuming. Here we present our approach for validating classification accuracy. Our experiments are designed to genotype from both sense and anti-sense directions, and we measure the concordance rate of results from both directions, which can be used to indicate the accuracy. The concordance rate is defined as

$$R_{concord} = |A \cap B| / |A|,$$

in which *A* and *B* represent the genotypes from sense and anti-sense directions, respectively. In other words, if the detected genotypes are the same for both directions, we say they are correct. Fig. 3 shows the blinded testing results using the canonical classifier as shown in Fig. 2. The concordance rate is 98%.

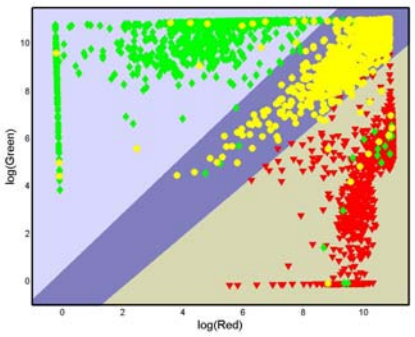

Fig. 3: Blinded testing results using GenoIterSVM.

5. Discussions and Future Work

When the number of heterozygous SNPs is too small (only a couple hundreds, for example), data renormalization could be strongly biased in the iterative genotype-calling procedure because the linear curve fit of points from a very small population may significantly distort the real picture of the difference in the two channels. This can be avoided by processing the data collectively from several individual microarray assays instead of doing single data set each time. This is the actual strategy we adopt in our investigation of the loss-of-heterozygosity (LOH) pattern in breast cancers.

Further investigation of the classification errors in Fig. 3 also shows that there are approximately 60 data points that are wrongly classified which are far from the decision boundaries, *i.e.*, the two straight lines separating the whole region into three distinct sub regions. Several factors may contribute to these mistakes. First, it is possible that during the microarray assays, amplification in the PCR and the hybridization on the glass substrate associated with these grid points are somewhat biased. Second, data points from the two directions of the DNA sequences lead to different genotype calls due allelic diversity at each DNA locus. If we could exclude those 60 problematic genotype calls in Fig. 4 in measuring the accuracy of the genotype-calling methods, the concordance rate reaches 99.4%. This is likely to be the accuracy limit for any classification scheme based on a data set that does not have replicas. Accuracy could further be improved by giving "N/D" (not determined) calls to data points that are within certain distance of the decision boundary. However, this will unavoidably increase the "NO CALL" rate, which is the dual of the accuracy (concordance). Accuracy could also be improved by having certain number of replicas in the microarray design.

Observations made here are essential for standardization in genotyping microarray experiments. For example, a minimum number of replicates should be included and genotypes from both sense and anti-sense directions should always be considered in probe/primer design.

6. References

[1] L. Kruglyak, and D. A. Nickerson, "Variation is the spice of life," *Nat Genet* 27: 234-6, 2001.

[2] http://www.ncbi.nlm.nih.gov/SNP/

[3] V. D. Schmith, D. A. Campbell, S. Sehgal, W. H. Anderson, D. K. Burns, L. T. Middleton and A. D. Roses, "Pharmaco-genetics and disease genetics of complex diseases," *Cell Mol Life Sci*, 60: 1636-46, 2003.

[4] H.-Y. Wang, M. Luo, H. Li, "A genotyping system capable of simultaneously analyzing >1,000 single nucleotide polymorphisms in a haploid genome," *Genome Research*, 2005 (to appear).

[5] W.-M. Liu, X. Di, G. Yang, H. Matsuzaki, J. Huang, R. Mei, T. B. Ryder, T. A. Webster, S. Dong, G. Liu, K. W. Jones, G. C. Kennedy and D. Kulp, "Algorithms for large-scale genotyping microarrays," *Bioinformatics*, 19(18), 2397-2403, 2003.

[6] C. Cortes, V. Vapnik, "Support vector network," *Machine Learning* 20: 273-297, 1995.

[7] G. C. Tseng, M-K. Oh, L. Rohlin, J. C. Liao and W. H Wong, "Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects," *Nucleic Acids Res*, 29(12): 2549-2557, 2001.

[8] Y. Xiao, C. Hunt, M. Segal and H. Yang, "A novel stepwise normalization method for two-channel cDNA microarrays," *Proceedings of the 26th Annual International Conference of the IEEE EMBS, 2921-2924, 2004.*