中央研究院
資訊科學研究所
Institute of Information Science, Academia Sinica • Taipei, Taiwan, ROC

TR-IIS-05-024

An Adaptive Prototype Classification Method with Applications to Genetic Marker Selection

Ke-Shiuan Lynn, Chin-Chin Lin, Wen-Harn Pan, and Fu Chang

December 2005 || Technical Report No. TR-IIS-05-024
http://www.iis.sinica.edu.tw/LIB/TechReport/tr2005/tr05.html

An Adaptive Prototype Classification Method with Applications to Genetic Marker Selection

Ke-Shiuan Lynn[1], Chin-Chin Lin[2], Wen-Harn Pan[1], and Fu Chang[3]
[1]Institute of Biomedical Sciences, Academia Sinica, Taiwan,
[2]Department of Electrical Engineering, National Taipei University of Technology, Taipei, Taiwan
[3]Institute of Information Science, Academia Sinica, Taipei, Taiwan

Abstract

Motivation: Ethnic origin is a complex trait that can be affected via multiple genetic factors. The traditional method, based on studying one gene or a few genes at a time, is not effective in profiling such a complex nature. Due to the advancement in high throughput genotyping, massive polymorphism (marker) information becomes available. Polymorphisms contain information on individuals' inherited traits including disease susceptibility, physical appearance, ethnic origins, etc. However, typing multiple genetic markers can still be costly, and constructing an appropriate ethnic classifier may involve heavy computation. To cope with these problems, we propose a new method that can accomplish two things at a low computational cost: finding a minimum number of genetic markers and constructing an ethnic classifier based on this minimum set of markers.

Results: We present the following three types of results: (1) By testing on artificial datasets with specified degrees of separation, our results suggest that, when population groups have distinguished ethnic origins, the number of prototypes and the test accuracy of the classifier constructed by APL are nearly constant with respect to *n*, as long as *n* exceeds a threshold. On the other hand, when the groups are of high admixture, both the number of prototypes and the test accuracy of the constructed classifier become unstable. (2) The proposed adaptive prototype learning (APL) method has much lower training cost and comparable test accuracy to two other methods, STRUCTURE and Support Vector Machines (SVM). (3) In the largest dataset consisting of 661 individuals, we are able to achieve 98.8% accuracy at top-36 markers chosen from 431 STRP markers, and 99.4% accuracy at top-48 markers chosen from the same set 431 markers. This is a rather favorable result in comparison with two former studies that achieve lower accuracy rates at higher number of markers.

Availability: The algorithm presented in this paper has been implemented in C. Source code is freely available for download at:

http://dar.iis.sinica.edu.tw/Download%20area/apl.htm.

Contact: fchang@iis.sinica.edu.tw

1. Introduction

Accurate population stratification is essential in genetic researches and in medical decision-making processes. As indicated by several papers, however, ethnic origins derived from subjective self-reports can seriously bias the results of genetic studies (Helgason et al., 2005, Kittles et al., 2002, Pritchard et al., 2001, Risch et al., 2002). A more scientific and accurate alternative for ethnic identification can be performed by way of genetic markers. However, previous studies based on relatively small sets of genetic markers also drew contradictory conclusions on the usefulness of discrete genetic categories toward biomedical studies (Burchard et al., 2003; Cooper et al., 2003; Haga et al., 2003; Risch et al., 2002; Schwartz, 2001). With recent advances in unraveling the human genome, profiling of ethnic origin using multiple genetic polymorphisms has become possible (Hoggart et al, 2004, Patterson et al., 2004, Pritchard et al., 2000). In addition, recent applications of some cluster analysis methods to large sets of genetic markers demonstrated that the resultant clusters exhibit good concordance with the commonly used ethnic groups: African, European/West Asian, East Asian, Pacific Islanders, and Native American (Bowcock et al., 1994; Calafell et al., 1998; Rosenberg et al., 2002; Tang et al., 2005). It was pointed out that the previous contradictory conclusions could be resulted from the number and the type of genetic markers used in the studies (Risch et al., 2002).

Nevertheless, genotyping a large number of markers can be costly. The effective representation of population stratification should rely on *the right markers*, rather than simply on a large number of markers. In this paper, we propose a method that can accomplish the following two things: finding the right set of markers and constructing a classier on this set. The proposed method consists of two technical ingredients. First, when a set of markers is given, an adaptive prototype learning (APL) algorithm is used to determine the location and the number of prototypes (cluster centers) for each ethnic group. Second, the set of markers is ranked based on a metric, called information gain, and APL is applied to the sets with an increasing number of ranked markers, until a desirable classification error is reached.

We first demonstrate the effectiveness of the classifier constructed by our method using artificial datasets with specified degrees of separation. We then apply the proposed method to four datasets consisting of various subpopulations excerpted from HGDP-CEPH Human Genome Diversity Cell Line Panel (Cann et al., 2002). To compare APL with some existing classification algorithms, we also apply the following algorithms to the same datasets: a model-based approach adopted by STRUCTURE (Pritchard et al., 2000), and a margin-based approach adopted by SVM (Vapnik, 1995). The results show that our method achieves results at comparable accuracies with other two methods, but at a much faster training speed.

In the largest dataset that comprises 127 Africans, 108 Southern Americans, 226 East Asians, 161 Europeans, and 39 Oceanians, totaling to 661 individuals, our method achieves 98.8% accuracy at top-36 markers chosen from 431 STRP markers, and 99.4% accuracy at top-48 markers chosen from the same 431 markers. These results are favorable as compared with some previous studies, based on smaller datasets, that achieve lower accuracy rates at higher number of markers.

2. The Method

2.1 Construction of Feature Vectors out of Genetic Markers

With the exception of sex chromosomes, each human chromosome has two copies, one from the male and the other from the female parent. For convenience of discussion, we denote one of the two copies as *u*-copy and the other as *v*-copy. To label a copy as *u* or *v* has no implication of its parental origin, since this information is not available from the chromosome itself. We also assume that there are *L* genetic markers that we want to study. Each marker has also two copies, each residing in a chromosome copy. For the l^{th} marker, the data we collect from its two copies are denoted as u_i and v_i, $l = 1, 2, \ldots, L$. We thus form two vectors out of the *L* markers:

$$\begin{cases} \mathbf{U} = \{u_1, u_2, \ldots, u_L\}, \\ \mathbf{V} = \{v_1, v_2, \ldots, v_L\}. \end{cases} \tag{1}$$

To compare the genetic polymorphisms between two individuals, we can compare their **U** and **V**. For this purpose, we first derive a feature vector out the two vectors. In case u_i and v_i assume only two values: 1 if a certain feature has been found and 0 otherwise, we obtain the feature vector as

$$\mathrm{F}_1(\mathbf{U}, \mathbf{V}) = \mathbf{U} + \mathbf{V}. \tag{2}$$

This derivation of feature vectors is suitable for *Alu* insertion markers, which are genetic markers consisting of the *presence/absence* of “a family of non-coding DNA sequence” (Makalowski, 1995). Such a derivation, however, is not adequate for short tandem repeat polymorphism (STRP) markers (Weber and May, 1989). In the latter context, where u_i and v_i denote the number of times an STRP marker repeats itself, these numbers should be understood as indices rather than numbers in a coordinate system. Thus, to obtain a feature vector for them, we have to first transform the multi-valued **U** into a binary-valued $\mathbf{U}^b$, where the l^{th} component of **U** expands to N_l components of $\mathbf{U}^b$ and N_l is the number of possible values of u_i. We perform a similar transformation of **V** into $\mathbf{V}^b$, and then obtain the feature vector as

$$\mathrm{F}_2(\mathbf{U}, \mathbf{V}) = \mathbf{U}^b + \mathbf{V}^b. \tag{3}$$

where the expanded feature vector F_2 has the length of $D = \sum_{l=1}^{L} N_l$. Thus, for example, if $\mathbf{U} = (1, 2)$, $\mathbf{V} = (3, 2)$, $N_1 = 3$, and $N_2 = 4$, we obtain $\mathbf{U}^b = (1, 0, 0, 0, 1, 0, 0)$, $\mathbf{V}^b = (0, 0, 1, 0, 1, 0, 0)$, and $\mathrm{F}_2(\mathbf{U}, \mathbf{V}) = (1, 0, 1, 0, 2, 0, 0)$. In fact, from $\mathrm{F}_2(\mathbf{U}, \mathbf{V})$ and the fact that $N_1 = 3$, and $N_2 = 4$, we immediately infer that one copy of the first marker does not repeat itself and another copy has 2 repetitions, while both copies of the second marker have two repetitions.

When we have obtained a feature vector out of each gene profile, we are able to compute the L_2-distance between two feature vectors $\mathbf{F} = (f_1, f_2, \ldots, f_L)$ and $\mathbf{G} = (g_1, g_2, \ldots, g_L)$, defined as

$$\| \mathbf{F} - \mathbf{G} \|^2 = \sum_{l=1}^{L} (f_l - g_l)^2 . \tag{4}$$

2.2 Adaptive Prototype Learning Algorithm

A prototype classification method, which has been successfully applied to the recognition of character images (Chang et al., 2004a; Chang et al., 2004b; Chang et al. 2005; Chou et al., to appear; Liu et al., 2005) and potentially applicable in many other multi-class classification problems, is used in this paper for classifying human population via their genetic polymorphisms. The proposed method is basically a clustering method. However, unlike the approach in Rosenberg et al., 2002, which forms clusters from samples of different labels (ethnic types), our method forms *homogenous* clusters in the sense that each cluster consists of samples of the same label. Our method thus starts with a set of training samples labeled with their ethnicity. A learning algorithm then proceeds to determine the number as well as the location of prototypes, whereas prototypes are defined as prototypes. The algorithm consists of two loops. The *outer* loop decides whether all clusters are homogeneous and, when they are not, identify those labels for which we want to build more prototypes. The *inner* loop computes the prototype locations for the number of prototypes specified by the outer loop.

We use the fuzzy c-means (FCM) clustering technique (Bezdek, 1981) in the inner loop to compute the prototype locations. FCM assigns, to each sample $\mathbf{x}$ and a given cluster center C, a grade of membership that varies inversely with the distance between $\mathbf{x}$ and C. The cluster center is the weighted average of all samples, with grades of membership serving as the weights. This technique relies on an iterative process to find the location of cluster centers. The convergence of the iterative process is always guaranteed, although not necessarily to the optimal value of the squared error criterion.

In the problem of ethnic classification, we assume that all individuals are presented as feature vectors (see Section 2.1) in the *D*-dimensional Euclidean space. Prototypes are also vectors in the same space and do not have to be samples per se. The prototype construction process is designed to determine the number and location of prototypes. When FCM is used to adjust the location of prototypes, there is no guarantee that all samples get absorbed eventually (Chang et al., 2004a). In order to ensure that the construction process terminates, we have to make a special control as follows. If an unabsorbed sample $\mathbf{x}$ is used as a seed for generating new *C*-prototypes, we check whether this addition produces any empty domain of attraction (DOA), where the DOA of a prototype $\mathbf{p}$ is defined as the set of samples of the *same* label that find $\mathbf{p}$ as the nearest prototype. If the DOAs are empty, we declare $\mathbf{x}$ as *futile* and restore all old *C*-prototypes. When a sample is declared futile at some iteration, it will not be taken as a sample at any later iteration. The process terminates, when all samples are either absorbed or declared as futile, whereas a sample $\mathbf{x}$ is *absorbed* if there is a prototype $\mathbf{p}$ such that $\mathbf{p}$ has the same label as $\mathbf{x}$ and $||\mathbf{x} - \mathbf{p}||^2 < ||\mathbf{x} - \mathbf{q}||^2$ for all other prototypes $\mathbf{q}$. The details of this process, referred to as adaptive prototype learning or APL, are stated as follows.

```
1. For each label $C$, initiate a $C$-prototype and set $n(C) = 1$.
2. Set all_absorbed = 0.
3. Outer loop: while all_absorbed equals 0 {
4.         For each sample, perform absorption check.
5.      If no more un-absorbed samples exists, set $all\_absorbed$ = 1.
6.      If there are un-absorbed samples
7.          **Inner loop:** for each label $C$ {
8.              If there are still unabsorbed $C$-samples
9.                  Set $to\_augment$ = 1.
10.             Else
11.                 Set $to\_augment$ = 0.
12.             Endif
13.             While $to\_augment$ equals 1 {
14.                 Select a $C$-sample $\mathbf{x}$ out of unabsorbed $C$-samples.
15.                 Employ FCM to determine $n(C)$+1 $C$-prototypes, using $\mathbf{x}$
                        and all existing $C$-prototypes as seeds.
16.                 If some of the prototypes have empty DOA
17.                     Declare $\mathbf{x}$ as futile.
18.                 Else
19.                     set $to\_augment$ = 0.
20.                 Endif
21.             }
22.             $n(C) = n(C)+1$.
23.         **} End of inner loop**
24.     Endif
25. **} End of outer loop**
```

In Line 1, the initial C-prototype is set to be the statistical average of all C-samples. In Line 14, the selection of **x** is made in the following way. Let Ψ_C be the set of un-absorbed C-samples. Each member of Ψ_C casts a vote to the nearest member in this set. The member that receives the highest vote is selected as **x**. In Line 15, FCM is employed to update prototypes as follows. Let $\mathbf{x}_j$, j = 1, 2, …, J_C, be a set of C-samples. The center $\mathbf{c}_i$ of cluster i, i = 1, 2,…, $n(C)$+1, is determined as a weighted average of all samples by

$$\mathbf{c}_i = \frac{\sum_{j=1}^{J_C} u_{ij}^m \mathbf{x}_j}{\sum_{j=1}^{J_C} u_{ij}^m}, \tag{5}$$

where u_{ij} is the grade of membership of $\mathbf{x}_j$ in cluster i and m is the *fuzzifier parameter* specified by users. In all applications considered in this paper, the fuzzifier parameter m is set to 1.1. The grade of membership u_{ij} is determined by

$$u_{ij} = \frac{\left(1/\|\mathbf{c}_i - \mathbf{x}_j\|\right)^{\frac{2}{m-1}}}{\sum_{k=1}^{J_C} \left(1/\|\mathbf{c}_i - \mathbf{x}_k\|\right)^{\frac{2}{m-1}}} \quad . \tag{6}$$

FCM is an iterated procedure that first computes grades of membership using (6) with $\{\mathbf{c}_i\}$ being the set of seeds. It then updates the cluster centers using (5) and the grades of membership using (6). This process continues until the number of iterations reaches 50, or $\sum_i \|\mathbf{c}_i^{old} - \mathbf{c}_i^{new}\| = 0$. The final cluster centers are then assigned as new prototypes.

Proof for the convergence of the APL:

The number of futile samples is bounded from above, since it cannot exceed the total number N of samples. So we assume that the last futile sample is created at iteration J, with $J \leq N$. If all samples are absorbed at the end of J, we are done with the proof. Otherwise, we will continue to create more prototypes, all with non-empty DOAs. If the number of unabsorbed samples never deceases to zero, then we shall eventually have more DOAs than samples, implying that some DOA is empty, conflicting with the requirement that all DOAs are non-empty. □

What we have just described is the hard version of APL, which constructs as many prototypes as possible to accomplish a zero training error rate. This may not fit well with the situation in which noisy samples exist in the training dataset. To insist on a zero error rate in this case can compromise the prediction power of the resultant classifier. A better way to do is to maintain the error rate to a level so as to generate the best test accuracy. To do so requires a cross-validation task, in which we randomly divide the training data into K parts, or K folds. In each trial, we use K-1 folds as training data and construct prototypes from them with the APL, and the remaining fold as the validation data, on which the test accuracy the prototypes can be tested.

In the trial in which the k^{th} fold is used as the validation data, we construct prototypes and record the following information. For a given level of training error rate e, we record the lowest number $n_k(e)$ of iterations at which the training error rate falls below e. We also compute the validation accuracy $v_k(e)$ for the prototypes constructed at iteration $n_k(e)$. Let $v(e) = \sum_{k=1}^{K} v_k(e)/K$. The optimal training error rate is then

$$e_{opt} = \operatorname*{aug\,max}_{e} v(e). \tag{7}$$

At the end of the prototype construction process, i.e., when a zero training error rate is achieved, we are able to obtain all $v(e)$ for all e, and thus $v(e_{opt})$. This constitutes the soft version of APL. From this point on, we assume that the soft version of APL is used.

2.3 Selection of Genetic Markers

The prototypes built by APL for ethnic groups turns out to be quite stable, as reflected in the following fact. Suppose that the total quantity of genetic markers is M. We first rank these M markers according to a metric, to be defined in a moment. We then sort these markers according to this metric. Let F_n be the set of top-n markers, $n = 1, 2, \ldots, M$. We build prototypes on the feature vectors whose features are taken from F_n. We denote the number of prototypes as P_n. It turns out that P_n stays nearly constant when we increase n by one at a time. The relative stability of P_n means that a subset of genetic markers is can be as good as the whole set of marker for discriminating ethnic types. It also means that researchers can focus their efforts on such a subset and achieve the same results as working with a gigantic set.

To be able to derive a reduced set of markers, we must provide a measure of discriminative power for markers. Note that the concern here is *which* and *how many* genetic markers are sufficient. This is not the same as what is dealt in dimension reduction, whose goal is to find a reduced set of linearly combined features. What we are interested here is a subset of markers, rather than some linear combinations of all markers.

Let us first define a few terms. $\mathbf{C} = \{C_1, C_2, \ldots, C_i, \ldots\}$ be the collection of ethnic types, and $\{m_1, m_2, \ldots, m_j, \ldots\}$ be all possible values that can be obtained from marker *m*. One problem associated with the marker, as we mentioned before, is that a marker has two copies. Since they may assume two different values, we view each copy as a separate sample. The two copies of the same marker are therefore viewed as two independent samples. The discriminative power of a marker is then defined as the gain of information:

$$\begin{aligned} H(\mathbf{C})-H(\mathbf{C}\,|\,m) &= -p(\mathbf{C})\log(\mathbf{C})-\left[-\,p(\mathbf{C}\,|\,m)\log(\mathbf{C}\,|\,m)\right] \\ &= -\sum_i p(C_i)\log p(C_i)+\sum_j p(m_j)\sum_i p(C_i\,|\,m_j)\log p(C_i\,|\,m_j) \qquad (7) \\ &= -\sum_i \frac{|C_i|}{|C|}\log\frac{|C_i|}{|C|}+\sum_j \frac{|m_j|}{|m|}\sum_i \frac{|C_i \wedge m_j|}{|m_j|}\log\frac{|C_i \wedge m_j|}{|m_j|}, \end{aligned}$$

where $|C_i|$ = the number of samples whose ethnic type is C_i, $|C|=\sum_i |C_i|$, $|m_j|$ = the number of samples whose *m*-marker assumes value m_j, $|m|=\sum_j |m_i|$, and $|C_i \wedge m_j|$ = the number of samples whose ethnic type is C_i and whose marker *m* assumes value m_j.

The information gain is the difference between the entropy $H(\mathbf{C})$ and the conditional entropy $H(\mathbf{C}\,|\,m)$. It measures how much uncertainty about ethnic identities can be reduced due to the information carried by the marker *m*. We rank all markers according to this metric. For n = 1, 2, …, M, let P_n be the number of prototypes for top-*n* markers, and v_n the validation accuracy of these prototypes. Note that both P_n and v_n are average numbers, since they are averages of results obtained in *K* trials. The minimal number of markers is then chosen to be the smallest *m* such that $|v_n - v_M| \le \varepsilon$ and $P_n \le P_M$ for $m \le n \le m + n_0$, where ε is a positive real number and n_0 a positive integer, specified by users. In Section 3, we shall specify the values of ε and n_0 for the datasets on which we perform the experiments.

3. Experiment Results.

In this section, we first apply APL to some artificial datasets, each of which consists of two ethnic groups with a specified degree of separation. We then apply APL to some real ethnic datasets and also to compare its performance with that of two other algorithms: STRUCTURE, and SVM. In both experiments, we employ 5-fold cross-validation. The average accuracy on the validation data is denoted as the validation accuracy.

3.1 Performance Evaluation on Artificial Data

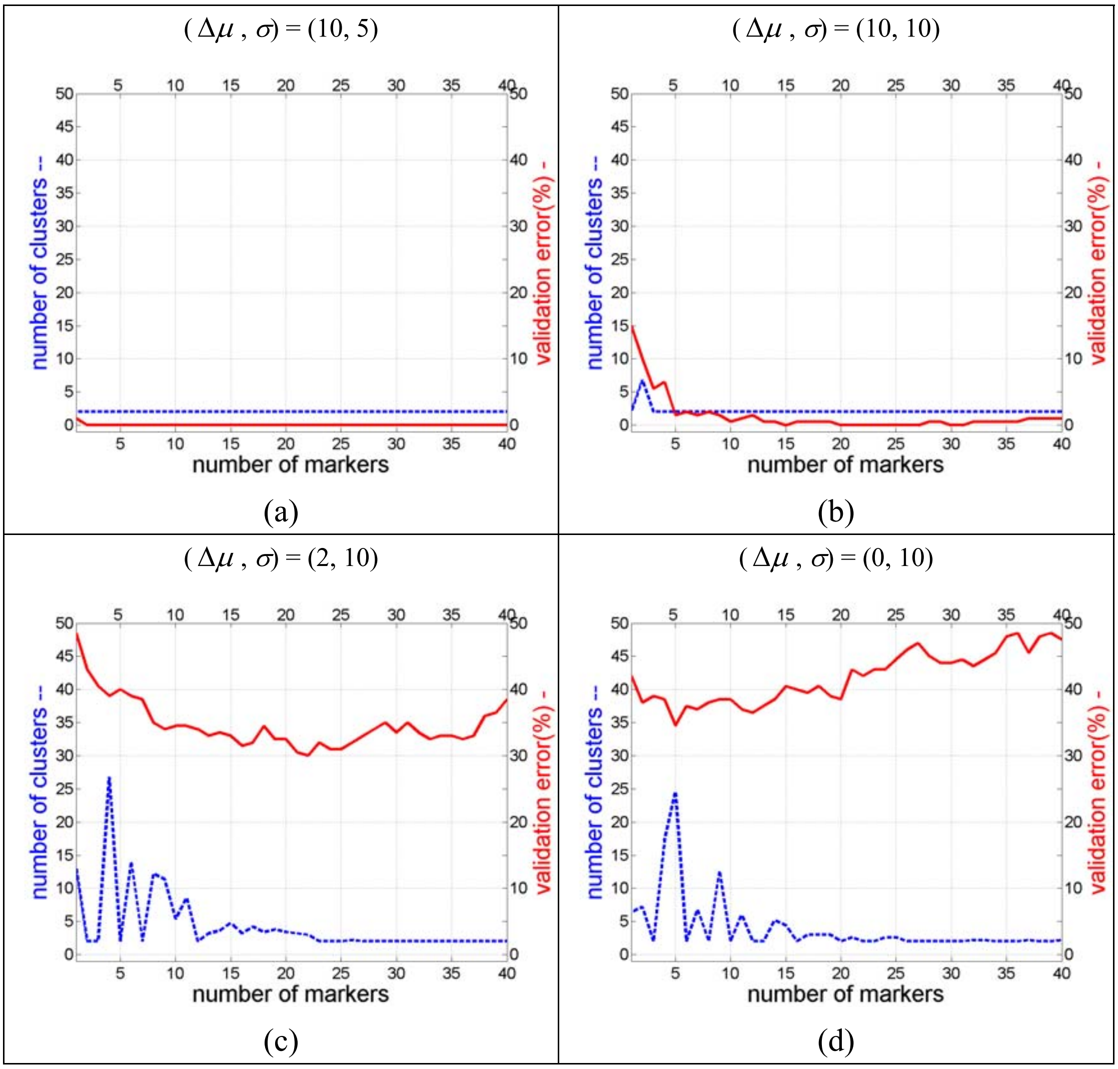
Before testing on real datasets, we wish to investigate the effectiveness of APL with some artificial datasets. For this purpose, we generate a dataset that has two ethnic groups G_1 and G_2, each of which consists of n individuals. We assume that each individual has two copies of chromosomes and d genes. Thus, for ethnic group g, we generate $2n$ vectors of the form

$$(x_{i1}^g, x_{i2}^g, ..., x_{id}^g) \tag{8}$$

where x_{ij}^g is a random variable, g = 1, 2; i = 1, 2, …, 2n; j = 1, 2, …, d. We further assume that two consecutive vectors of the form (8) represent two copies of chromosomes of the same individual.

The degree of separation between G_1 and G_2 is controlled by parameters $\Delta\mu$ and σ, where $\Delta\mu$ determines the mean differences between G_1 and G_2 and σ determines the concentration of each group (a low σ corresponds to a high concentration). In our experiment, we consider four pairs of ($\Delta\mu$, σ): (10, 5), (10, 10), (2, 10), and (0, 10). In addition, we set $n = 100$ and $d = 40$.

Figures 1a-1d correspond to the four pairs of ($\Delta\mu$, σ). Each of these figures contains two curves that plot the number of prototypes and the validation error (= 100% - validation accuracy) of these prototypes for a given number of markers. If the number of markers is *n*, it is understood that markers of top-*n* ranks are employed. Thus, in Figure 1a, the two ethnic groups are highly separated, since the mean difference between them is large ($\Delta\mu$=10) and each group is rather concentrated (σ=5). The solid line shows that APL builds one prototype per ethnic group, when the number of markers exceeds 2. The dashed line shows that APL achieves 0% validation error, under the same condition. As we move on to the other three pairs of ($\Delta\mu$, σ), the two ethnic groups become less and less separated, since their mean differences become smaller and each group is less concentrated, so APL builds more and more prototypes for the same number of markers, as shown in Figures 1b–1d. However, even when the two ethnic groups are seriously overlapped, as in the case of Figure 1d, the average number of APL prototypes per ethnic group is close to 1, when the number of markers exceeds 25.



$(\Delta\mu, \sigma) = (10, 5)$ (a) $(\Delta\mu, \sigma) = (10, 10)$ (b) $(\Delta\mu, \sigma) = (2, 10)$ (c) $(\Delta\mu, \sigma) = (0, 10)$ (d)

Figure 1. The clustering performances of APL during the marker-selection process evaluated using six datasets of different degrees of separations.

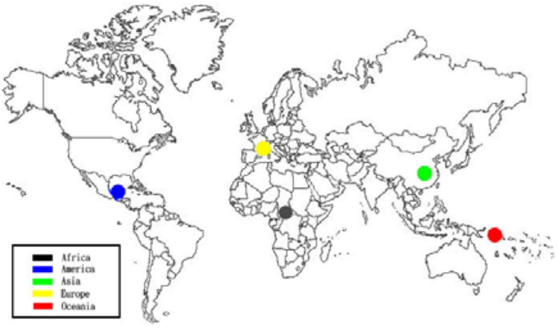
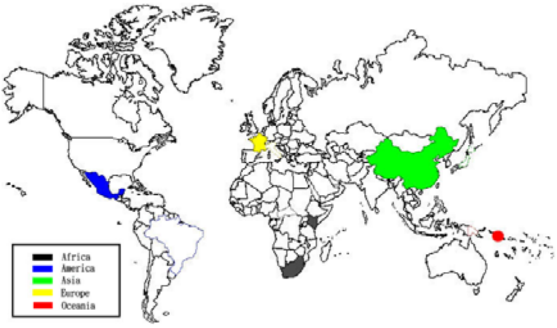
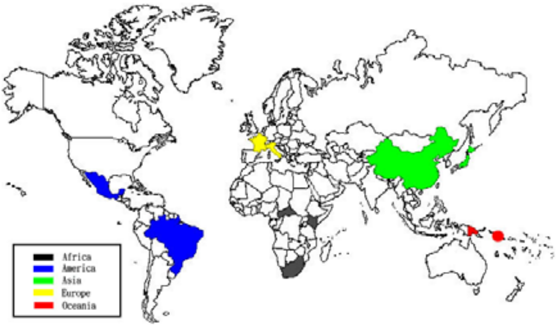
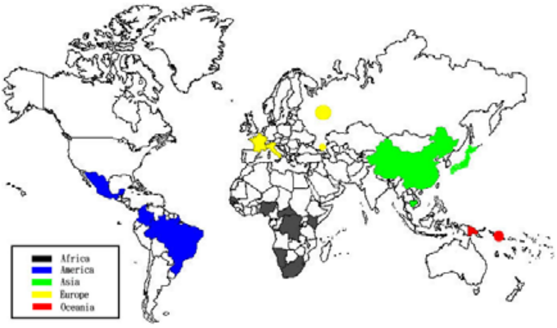
3.2 Performance Evaluation on Real Ethnic Data

To create datasets for the training and testing purpose, we draw our data from the database HGDP-CEPH Human Genome Diversity Cell Line Panel (Cann et al., 2002), in which a total of 1,064 lymphoblastoid cell lines (LCLs) from 1,051 individuals in 51 different subpopulations were collected, and 431 STRP markers were typed from the DNA of the LCLs. We create four datasets, denoted as A, B, C, and D, each of which consists of genotyping data of various subpopulations, excerpted from five major continents: Africa, America, Asia, Europe, and Oceania. As shown in Table 1, the population per continent in the four datasets is composed of individuals from a single race, a single nation, two nations, and multiple nations, respectively.

When applying APL to these datasets to produce prototypes for a given number of markers, we also apply SVM and STRUCTURE to the same data. The three algorithms, APL, SVM, and STRUCTURE, are all conducted in a 5-fold cross-validation, so that each of these algorithms produces its own validation accuracy for a given number of markers. For SVM, we employ RBF-based soft-margin version, where the value range of the RBF parameter γ is taken as $\{10^a$: a = -6, -5, …, 2$\}$, and the value range of the penalty factor C is taken as $\{10^b$: b = -4, -2, …, 3$\}$. The search for the optimal γ and C is through the 5-fold cross validation. Figures 2 plots the accuracies achieved by the three algorithms for the four datasets. The STRUCTURE result for dataset D is missing, however, due to the untraceable failure of the executable code supplied by Pritchard (Pritchard et al., 2000).

Table 1. The subpopulations excerpted from each continent.

Data distribution in dataset A	African	Biaka Pygmies (36)

	American	Maya (25)
	Asian	Han (45)
	European	French (29)
	Oceanian	NAN Melaneian (22)
Data distribution in dataset B 	African	Kenya (20)
	American	Mexico (50)
	Asian	China (184)
	European	France (53)
	Oceanian	Bougainville (22)
Data distribution in dataset C 	African	Kenya, Congo (56)
	American	Mexico, Brazil (95)
	Asian	China, Japan (215)
	European	France, Italy (103)
	Oceanian	Bougainville, NewGuinea (39)
Data distribution in dataset D 	African	Central African Republic, Democratic Republic of Congo, Senegal, Nigeria, Namibia, Kenya (127)
	American	Mexico, Brazil, Colombia(108)
	Asian	China, Japan, Cambodia (226)
	European	France, Italy, Orkney Islands, Russia Caucasus, Russia (161)
	Oceanian	Bougainville, NewGuinea (39)

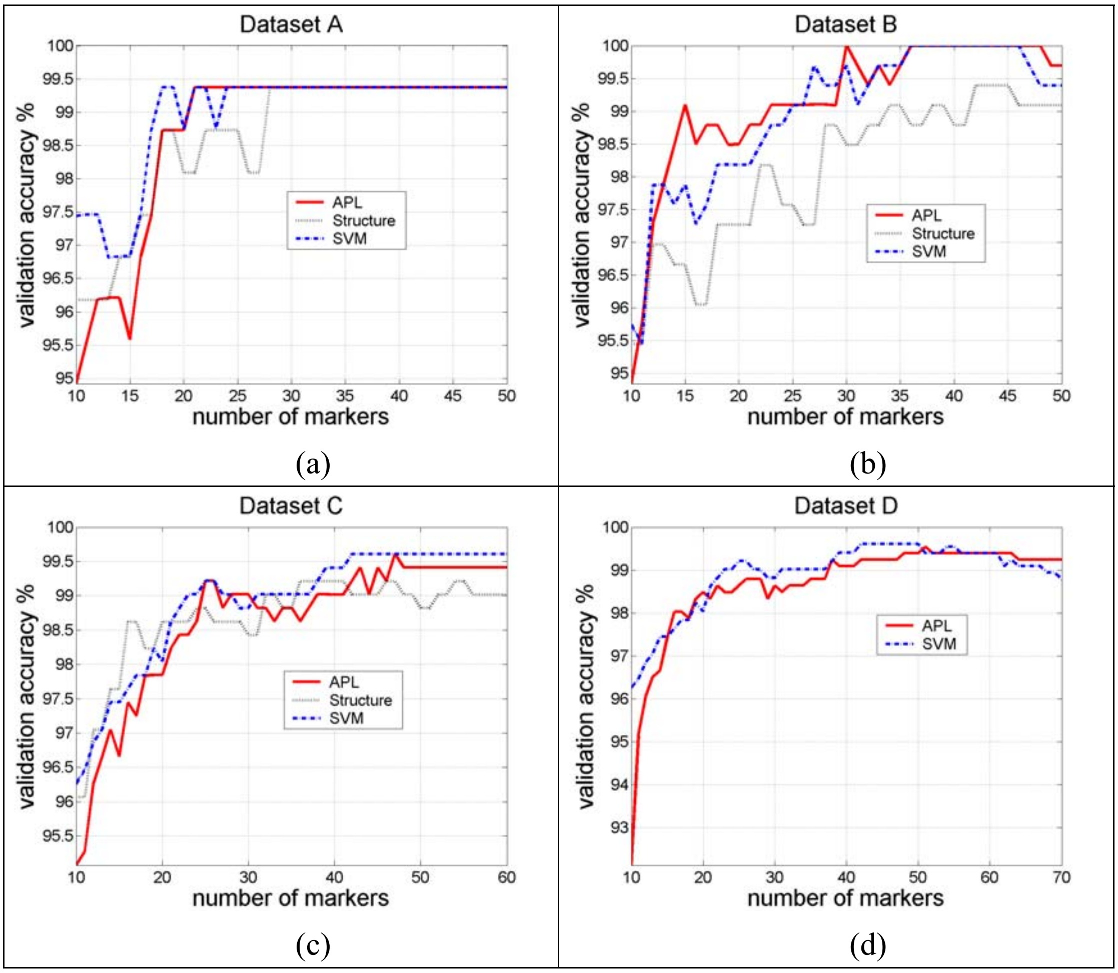


Figure 2. Comparison of validation accuracies achieved by APL, STRUCTURE, and SVM applied to the four datasets.

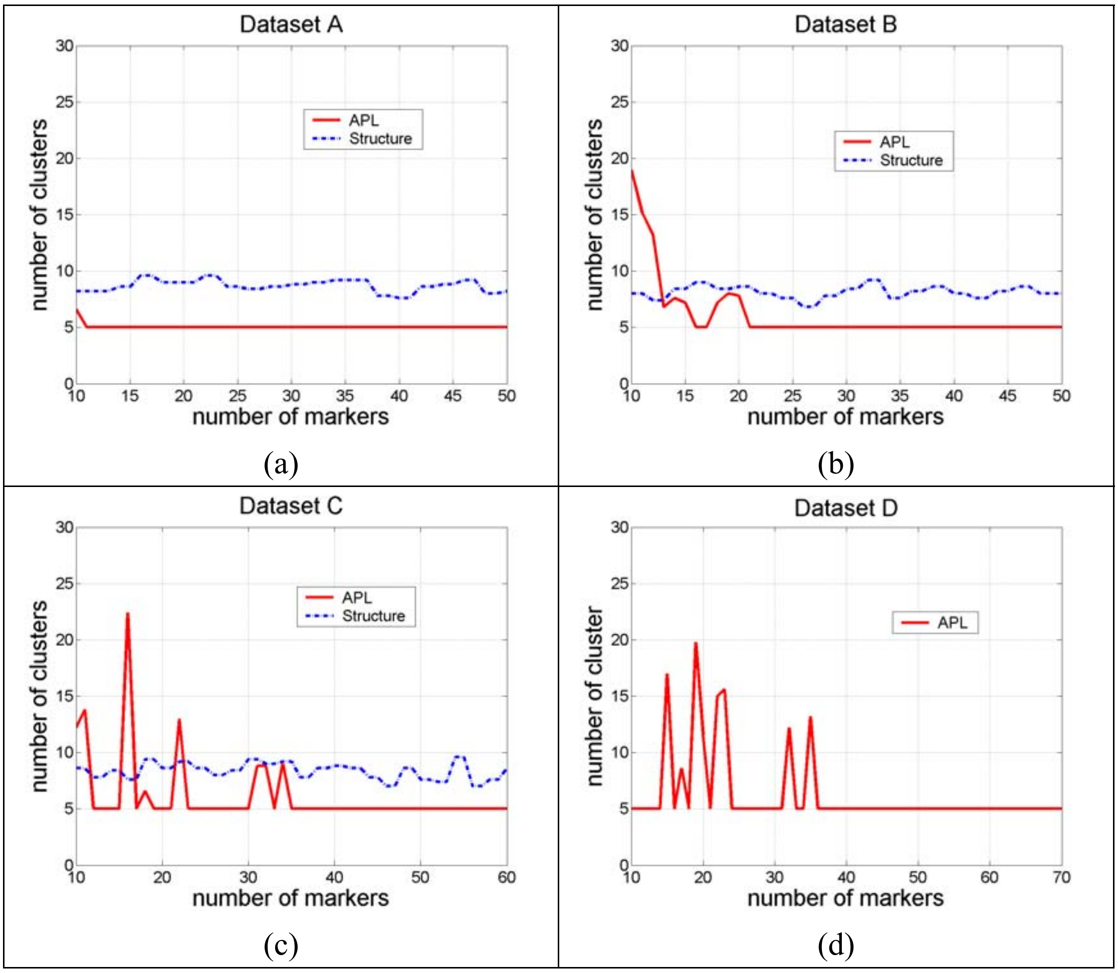
It is shown in Figure 2 that the three algorithms achieve comparable accuracies (the differences are mostly within 1%) in the four datasets and the accuracies degenerate in a similar fashion as the number of excerpted subpopulations per continent increases. Despite of the similarity in their results, the APL is advantageous in its much shorter training time compared with those of STRUCTURE and SVM, as shown in Table 2. It is seen there that APL spends seconds to accomplish the job, while STRUCTURE and SVM usually have to spend hours.

Table 2. Comparison of training time by APL, STRUCTURE, and SVM, applied to

the four datasets.

Dataset	APL	STRUCTURE	SVM
A	4 sec	2hr 23 min 8 sec	31min 28 sec
B	10 sec	4hr 12 min 58sec	1hr 14min 20sec
C	19 sec	6hr 56min 25sec	3hr 0min 47sec
D	29 sec	---	5hr 7min 31sec

In Section 2.3, we define the minimal number of markers to be m such that $P_n \le P_M$ and $|v_n - v_M| \le \varepsilon$ for $m \le n \le m + n_0$. Throughout all the experiments, we set ε = 0.005 and n_0= 10. Recall that the APL prototypes are cluster centers. It is therefore interesting to compare the number of clusters obtained by APL with those by STRUCTURE, which is also a clustering-based method. The results are shown in Figure 3. In all the four datasets, APL obtains five prototypes (one prototype per continent), for $m \le n \le m + 10$. The minimal numbers of markers are found to be 21, 35, 38, and 36 respectively. As shown in Figures 3a to 3c, the number of APL prototypes is much more stable than the number of STRUCTURE clusters. In Figure 3d, we do not plot the number of STRUCTURE clusters, due to the failure of the executable code on the dataset D.



(a)

(b)

(c)

(d)

Figure 3. Number of clusters obtained by APL and by STRUCTURE.

Dataset D is the largest dataset among the four, which comprises 127 Africans, 108 Southern Americans, 226 East Asians, 161 Europeans, and 39 Oceanians, totaling to 661 individuals. Our method, using APL as the classification technique and information-gain as the as the ranking metric, achieves 98.8% at top-36 markers, and 99.4% at top-48 markers, as listed in Table 3. In comparison, we quote two previous results by Bamshad et al., 2003, and by TuraKulov and Easteal, 2003. Both methods use STRUCTURE as the classification technique and randomization as the means for ranking, while using different types of markers to be described in a moment. Bamshad et al. works on a dataset consisting of 58 Africans, 67 Asians, and 81 Europeans, totaling to 206 individuals. Their method achieves 90% accuracy at 60 markers, randomly chosen from 160 markers (consisting of 100 *Alu* and 60 tetranucleotide microsatellites), and 99% accuracy at 100 markers randomly chosen from the same 160 markers. TuraKulov and Easteal, on the other hand, work on a dataset consisting of 30 Afro-Americans, 30 Asians, and 30 Caucasians, totaling to 90 individuals. Their method achieves 90% accuracy at 100 markers randomly chosen from 5,074 SNP markers.

Table 3. Top-48 STRP markers employed by our method, resulting in 99.4% accuracy on dataset D.

	STRP markers
Chromosome 1	GTTTT002P, TTTA063P, ATA43C09M, ATA20F08P, GATA2B02Z, AAT252, AAT258
Chromosome 2	GATA181G08M, AAT263P, ATA16D09
Chromosome 3	ATC3D09, ATA57D10M, AAC030
Chromosome 4	ATT077P, TAGA049, ATT015, AATA045, GATA150B10
Chromosome 5	GATA12G02
Chromosome 6	AGC001b, GATA30A08M, SE30, ATA1F08, TATC050zM
Chromosome 9	GATA61F04, ATA42G04P
Chromosome 11	AAT265M, AAT268
Chromosome 12	AAT262, ATA080M, ATA63A05P
Chromosome 13	GTT035
Chromosome 14	ATAC026P, ATGG002
Chromosome 16	TTTA028, AAT226, ATA063
Chromosome 17	AAT095, AAT083
Chromosome 18	CTG008
Chromosome 19	TTTA075P
Chromosome 20	GATA65E01, AAAT007
Chromosome 22	TTA015P, AGAT055Z, AGAT121P

Chromosome 23	AAT193, AAAT112P

4. Conclusion

In this paper, we propose a method that selects markers and builds prototypes on these markers. This method works in the following order. First, it uses the information-gain metric to sort all markers. Second, it employs APL to construct prototypes on top-*n* markers, with increasing *n*. Third, it determines the minimal number of markers, based on cross validation results. This method is compared favorably with two other approaches, using SVM or STRUCTURE as classification techniques and the same metric for ranking markers: the three approaches achieve comparable accuracy results, while our method runs at a much faster speed than the other two in training. The experimental results also show that our method comes up with rather stable number of prototypes as well as accuracy rates, as the number of markers exceeds the minimum one. These results are also favorable in comparison with some previous studies that achieve lower accuracy rates at larger numbers of markers, while working on smaller datasets.

References

[1] Bamshad, M.J. et al., (2003) Human population genetic structure and inference of group membership. *Am. J. Hum. Genet.*, **72**, 578-589.

[2] Bezdek, J.C., (1981) *Pattern Recognition with Fuzzy Objective Function Algorithms*. New York: Plenum.

[3] Bowcock, A.M. et al., (1994) High resolution of human evolutionary trees with poly-morphic microsatellites. *Nature*, **368**, 455–457.

[4] Burchard, E.G.., et al. (2003) The importance of race and ethnic background in bio-medical research and clinical practice. *N. Engl. J. Med.*, **348**, 1170-1175.

[5] Calafell, F. et al., (1998) Short tandem repeat polymorphism evolution in humans. *Eur. J. Hum. Genet.*, **6**, 38–49.

[6] Cann, H.M. et al., (2002) A Human Genome Diversity Cell Line Panel, *Science*, **296**, 261-262.

[7] Chang, F. et al., (2004a) A Prototype Classification Method and Its Application to Handwritten Character Recognition, *IEEE SMC*, 4738-4743, Hague.

[8] Chang, F. et al., (2004b) Applying A Hybrid Method to Handwritten Character Recognition, *Intern. Conf. Pattern Recognition 2004*, **2**, 529-532, Cambridge.

[9] Chang F. et al. (2005), Caption Analysis and Recognition for Building Video Indexing Systems, *ACM Multimedia Systems Journal*, **10**, 344-355.

[10] Chou C.-H. et al., A Prototype Classification Method and Its Use in a Hybrid Solution for Multiclass Pattern Recognition, to appear in Pattern Recognition.

[11] Cooper, R.S. et al (2003) Race and genomics. *N. Engl. J. Med.*, **348**, 1166-1170.

[12] Haga, S.B. and Venter, J.C. (2003) Genetics. FDA races in wrong direction. *Science*, **301**, 466.

[13] Helgason, A. et al. (2005) An Icelandic example of the impact of population structure on association studies. *Nature Genetics*, **37**, 90-95.

[14] Hoggart C.J. et al. (2004) Design and analysis of admixture mapping studies. *Am. J. Hum. Genet.*, **74**, 965-978.

[15] Kittles, R.A. et al, (2002) CYP3A4-V and prostate cancer in African Americans: causal or confounding association because of population stratification? *Hum Genet.*, **110**, 553-60.

[16] Liu, Y.-H. et al., (2005) Language Identification of Character Images Using Machine Learning Techniques, *Inter. Conf. Document Analysis and Recognition*, Seoul.

[17] Makalowski, W. (1995) SINEs as a genomic scrap yard: An essay on genomic evolution; in *TheImpact of Short Interspersed Elements (SINEs) on the Host Genome* (Maraia, R.J. & Austin, R.G.,eds.) 81-104, Landes Company.

[18] Patterson, N. et al., (2004) Method for high-density admixture mapping of disease genes. *Am. J. Hum. Genet.*, **74**, 979-1000.

[19] Pritchard, J.K. et al., (2000) Inference of population structure using multilocus geno-type data. *Genetics*, **155**, 945-959.

[20] Pritchard, J.K., Donnelly, P. (2001) Case-control studies of association in structured or admixed populations. *Theor. Pop. Biol.*, **60**, 227-237

[21] Risch, N., et al. (2002) Categorization of humans in biomedical research: genes, race and disease. *Genome Biol.*, **3**, comment2007.

[22] Rosenberg, N.A. et al., (2002) Genetic structure of human populations. *Science*, **298**, 2381–238.

[23] Schwartz, R. S. (2001) Racial profiling in medical research. *N. Engl. J. Med.*, **344**, 1392-1393.

[24] Tang, H, et al. (2005) Genetic structure, self-identified race/ethnicity, and confounding in case-control association studies. *Am. J. Hum. Genet.*, **76**, 268-275.

[25] Turakulov, R. and Easteal, S. (2003) Number of SNPS loci needed to detect population structure. *Human Heredity*, **55**, 37-45.

[26] Vapnik, V. (1995) *The Nature of Statistical Learning Theory*, Springer-Verlag, New York.

[27] Weber J.L. and May P.M. (1989) Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. *Am. J. Hum. Genet.*,

44, 388-396.