

Identification of Ancestry Informative Markers from Chromosome-Wide Single Nucleotide Polymorphisms Using Symmetrical Uncertainty Ranking

Theera Piroonratana*, Waranyu Wongseree*, Touchpong Usavanarong*, Anunchai Assawamakin†, Chanin Limwongse† and Nachol Chaiyaratana*†

*Department of Electrical Engineering, Faculty of Engineering

King Mongkut's University of Technology North Bangkok, Bangkok 10800, Thailand

Email: theepi@gmail.com, waranyu.wongseree@gmail.com, blood_serpent@hotmail.com, n.chaiyaratana@gmail.com

†Division of Molecular Genetics, Department of Research and Development, Faculty of Medicine Siriraj Hospital Mahidol University, Bangkok 10700, Thailand

Email: anunchai_ice@yahoo.com, siclw@mahidol.ac.th

Abstract—Ancestry informative markers (AIMs) have been proven to contain necessary information for population classification. In this article, round robin symmetrical uncertainty ranking for preliminary AIM screening is proposed. Each single nucleotide polymorphism (SNP) is assigned a rank based on its ability to separate two populations from each other. In a multi-population scenario, all possible population pairs are considered and the screened SNP set incorporates top-ranked SNPs from every pair-wise comparison. After the preliminary screening, SNPs are further screened by a wrapper which is embedded with a naive Bayes classifier. A classification model is subsequently constructed from the finally screened SNPs via a naive Bayes classifier. The application of the proposed procedure to the HapMap data indicates that AIM panels can be found on all chromosomes. Each panel consists of 11 to 24 SNPs and can be used to completely classify the CEU, CHB, JPT and YRI populations. Moreover, all panels are smaller than the AIM panels reported in previous studies.

Keywords—ancestry informative marker; attribute selection; HapMap; pattern recognition; single nucleotide polymorphism.

I. INTRODUCTION

Human evolution can be traced via various forms of genetic information [1], [2]. Many studies involving mitochondrial DNA (mtDNA) [1] and DNA variation on the Y chromosome [2] reveal that the present human species is originated from Africa. In fact, the migration of behaviourally modern humans from Africa to all continents takes place only approximately 70,000–50,000 years ago [3]. Through this course of migration, the population subdivision has occurred and has resulted in the emergence of new populations and ethnic groups.

With the presence of strong evidence that supports the occurrence of population subdivision, the genetic description of a population can be established. Furthermore, the clustering of individuals into many populations with different genetic backgrounds can be done automatically [4], [5], [6]. The task of assigning an unknown individual to the correct population can be carried out by inspecting his or her population-specific genetic patterns once the population

boundary is defined via genetics or self-reported ethnicity. These patterns usually consist of ancestry informative markers (AIMs)—genetic markers that exhibit substantially different allele frequencies between populations of descendants derived from mutually inbred ancestors. The identification of AIMs has been proven to be beneficial to many research areas including genetic epidemiology [7], [8] and forensic science [9].

The international HapMap project discovers over 3,000,000 single nucleotide polymorphisms (SNPs) in the genome of each human individual [10]. As a result, the search for SNP-based AIMs usually involves genome-wide SNP screening. Many measures including informativeness [11], t statistics [12], [13] and F statistics [13] have been proposed for SNP prioritisation. The screening is then carried out via a greedy search [11] or a ranking method [13]. Once the SNPs have been selected, their capability as AIMs can be validated via classification model construction. The classification task specifically involves the use of genotypic attributes from selected SNPs as inputs for identifying the ethnicity or population label of an individual. Standard machine learning techniques that have been successfully implemented as classifiers include a support vector machine [13] and genetic programming [14].

The genome-wide SNP screening indicates that AIMs spread across the whole genome [6], [12], [13]. In fact, only 14 SNPs are required for the complete classification between the CEU (Utah residents with northern and western European ancestry), YRI (Yoruba in Ibadan, Nigeria) and combined JPT (Japanese in Tokyo) and CHB (Han Chinese in Beijing) samples in the HapMap data [6]. Furthermore, the early works on AIM identification also suggest that a SNP set that is suitable for ancestry inference is not unique. However, complete classification between these four populations has never been achieved. In this article, a round robin symmetrical uncertainty ranking technique is proposed for SNP prioritisation and screening during the AIM search. The proposed method falls into the category

of filter-based approaches for attribute selection in machine learning [15]. The information-theoretic rank is calculated via a comparison of genotype distribution as defined by each SNP from two populations at a time. This makes the method suitable for the identification of AIMs among samples from two or more populations. Furthermore, the classification models which are constructed from screened SNPs by a naive Bayes classifier are also provided. The application of the proposed method to the HapMap data indicates that complete classification between four populations in the HapMap data is possible.

II. DATA SET

The data explored in this study is obtained from the public release #23a of HapMap data (Phase II, release date: March 2008), which is available in NCBI build 36 (dbSNP b126) coordinates. The data set consists of 3,619,209 SNPs in which the genotypic attribute value according to each SNP can be a homozygous wild-type, heterozygous or homozygous mutant genotype. These SNPs are extracted from 270 samples representing four populations: CEU, CHB, JPT and YRI. Both CEU and YRI data sets consist of 90 related samples—30 father-mother-offspring trios. In contrast, both CHB and JPT data sets contain 45 unrelated samples. Since the original HapMap data set is composed of related and unrelated samples, only 210 unrelated samples are considered. The sample reduction is carried out by removing offspring samples from both CEU and YRI data sets.

III. ALGORITHM

The proposed information-theoretic measure is based on symmetrical uncertainty [16]. Consider a classification problem that involves a sample set in which each sample is described by n_a discrete-valued attributes (SNPs) and a class (population) label. Let A be an attribute and C be the class. The entropy H of the class before and after observing the attribute is given by

$$H(C) = - \sum_{c \in C} p(c) \log_2 p(c) \quad (1)$$

and

$$H(C|A) = - \sum_{a \in A} p(a) \sum_{c \in C} p(c|a) \log_2 p(c|a), \quad (2)$$

respectively where p denotes the probability value as estimated from the sample set. The difference between the entropy of the class before and after observing the attribute is the information gain [17] which is given by

$$\begin{aligned} \text{Information Gain} &= H(C) - H(C|A) \\ &= H(A) - H(A|C) \\ &= H(A) + H(C) - H(A, C). \end{aligned} \quad (3)$$

The degree of correlation between the attribute and the class can subsequently be estimated via symmetrical uncertainty (SU) which is defined by

$$\begin{aligned} SU &= 2 \times \left[\frac{H(A) + H(C) - H(A, C)}{H(A) + H(C)} \right] \\ &= 2 \times \left[\frac{H(C) - H(C|A)}{H(A) + H(C)} \right]. \end{aligned} \quad (4)$$

It is noticeable that symmetrical uncertainty can be calculated from a quotient between the information gain and the sum of class entropy and attribute entropy. An attribute that has a high SU value is highly correlated with the class and is also an important attribute for classification. SU can be directly measured in classification problems with the number of classes greater than or equal to two. However, the calculation of SU for a common SNP from two populations at a time is more useful for the identification of AIMs. This is because AIMs can generally be identified when at least one new population is emerged from the ancestral population. For clarification, the measure is referred to as SU_2 when SU is evaluated to determine the suitability of using a SNP for classifying only two populations. In a multi-population problem, $\binom{n_p}{2} = n_p! / ((n_p - 2)!2!)$ SU_2 measures are required where n_p is the number of populations considered.

After the SU_2 values have been derived from all SNPs, a rank can be assigned to each SNP; high SU_2 values lead to high ranks. It is noticed that the rank which is based on an SU_2 value only reflects the importance of a SNP when the classification of two considered populations is concerned. The top n_r SNPs with the highest ranks are subsequently selected as screened attributes for the two-population classification. Again, $\binom{n_p}{2}$ sets of top-ranked SNPs can be extracted from the n_p -population data. The merging of top-ranked SNP sets is then carried out where the size of the merged SNP set is between n_r and $n_r \times \binom{n_p}{2}$.

IV. RESULTS AND DISCUSSION

The search for AIMs covers only SNPs from the same chromosome. Since the HapMap data contains a large amount of SNPs, the data set of interest is first partitioned into a number of smaller data sets. Each partition consists of 5,000 positionally consecutive SNPs except for the last partition from each chromosome, which is allowed to contain less than 5,000 SNPs. The round robin symmetrical uncertainty ranking (SU_2 ranking) described in the previous section is then applied to SNPs within each partition. The top-ranked SNPs from all partitions on the same chromosome are merged and further reduced by a wrapper embedded with a naive Bayes classifier and a best first search for attribute subset selection (NB-Wrapper). The classification model is subsequently constructed from the finally screened SNPs. Ten-fold cross-validation is applied during the experiment. The classification error is summarised in Table I. It can be

Table I
THE CLASSIFICATION ERROR OF A NAIVE BAYES CLASSIFIER WITH FINALLY SCREENED SNP INPUTS.

| Chr | Number of misclassified samples | | | |
|-----|---------------------------------|-------------|-------------|-------------|
| | $n_r = 50$ | $n_r = 100$ | $n_r = 200$ | $n_r = 300$ |
| 1 | 0 | 0 | 0 | 0 |
| 2 | 0 | 0 | 0 | 0 |
| 3 | 0 | 0 | 0 | 0 |
| 4 | 0 | 0 | 0 | 0 |
| 5 | 0 | 0 | 0 | 0 |
| 6 | 0 | 0 | 0 | 0 |
| 7 | 0 | 0 | 0 | 0 |
| 8 | 0 | 0 | 0 | 0 |
| 9 | 0 | 0 | 0 | 0 |
| 10 | 0 | 0 | 0 | 0 |
| 11 | 0 | 0 | 0 | 0 |
| 12 | 0 | 0 | 0 | 0 |
| 13 | 0 | 0 | 0 | 0 |
| 14 | 0 | 0 | 0 | 0 |
| 15 | 0 | 0 | 0 | 0 |
| 16 | 0 | 0 | 0 | 0 |
| 17 | 0 | 0 | 0 | 0 |
| 18 | 0 | 0 | 0 | 0 |
| 19 | 0 | 0 | 1 | 0 |
| 20 | 0 | 0 | 0 | 0 |
| 21 | 0 | 1 | 1 | 1 |
| 22 | 0 | 0 | 0 | 0 |

clearly seen that complete classification is possible in all chromosomes. Since complete classification can be attained regardless of the setting of n_r during the SU_2 ranking, $n_r = 50$ is proven to be sufficient.

The chromosome-wide search for AIMs reveals that a set of SNPs that can lead to complete classification is not unique. In fact, the number of SNPs required to achieve this level of classification accuracy is much smaller than those reported in early works by Park et al. [12], Paschou et al. [6] and Zhou and Wang [13]. A summary of the numbers of required SNPs from the early works and the present study is given in Table II. Park et al. [12] employ a nearest shrunken centroid method while Zhou and Wang [13] develop a modified t -test for SNP screening. Both approaches are filter-based attribute selection techniques where each SNP is prioritised by identifying its usefulness for separating all population classes from one another. This is different from the strategy embedded in the SU_2 ranking in which each SNP is prioritised according to its usefulness for separating classes in each class pair. This strategic difference is most likely to be the cause of the reduction in the sizes of AIM panels from those reported in the works by Park et al. [12] and Zhou and Wang [13]. Nonetheless, the strategy employed in the SU_2 ranking can be incorporated into both the nearest shrunken centroid method and the modified t -test. The modification should enhance the capability of both approaches, which could lead to the reduction in the sizes of AIM panels. In contrast to Park et al. [12] and Zhou and Wang [13], Paschou et al. [6] use a clustering technique to identify AIMs. In other words, the population labels are not considered during the SNP screening. As a result, larger

Table II
NUMBER OF SNPs REQUIRED FOR THE CLASSIFICATION OF HAPMAP DATA. THE THREE-POPULATION PROBLEM IS FORMULATED BY GROUPING JPT AND CHB SAMPLES INTO THE SAME CLASS.

| Reference | # Pop | Source | # SNPs | Accuracy (%) |
|---------------------------------|-------|--------|--------|--------------|
| Present study ($n_r = 50$) | 4 | Chr 1 | 14 | 100.00 |
| | | Chr 2 | 14 | 100.00 |
| | | Chr 3 | 11 | 100.00 |
| | | Chr 4 | 15 | 100.00 |
| | | Chr 5 | 19 | 100.00 |
| | | Chr 6 | 16 | 100.00 |
| | | Chr 7 | 15 | 100.00 |
| | | Chr 8 | 19 | 100.00 |
| | | Chr 9 | 16 | 100.00 |
| | | Chr 10 | 11 | 100.00 |
| | | Chr 11 | 16 | 100.00 |
| | | Chr 12 | 19 | 100.00 |
| | | Chr 13 | 19 | 100.00 |
| | | Chr 14 | 24 | 100.00 |
| | | Chr 15 | 13 | 100.00 |
| | | Chr 16 | 15 | 100.00 |
| | | Chr 17 | 20 | 100.00 |
| | | Chr 18 | 15 | 100.00 |
| | | Chr 19 | 15 | 100.00 |
| | | Chr 20 | 19 | 100.00 |
| | | Chr 21 | 18 | 100.00 |
| | | Chr 22 | 19 | 100.00 |
| Park et al. [12] | 3 | Genome | 82 | 100.00 |
| Zhou and Wang [13] | 3 | Genome | 64 | 100.00 |
| | 4 | Genome | 100 | 90.00 |
| Paschou et al. [6] | 3 | Genome | 14 | 100.00 |
| | 4 | Genome | 164 | 99.52 |
| | 4 | Genome | 64 | 98.57 |

AIM panels than those from the present study are selected to achieve the maximum distances between population clusters. Although the technique proposed by Paschou et al. [6] may be less effective in the case of HapMap data, the technique is highly effective when the population labels are not known a priori and the population boundary is determined solely via genetics.

V. CONCLUSION

In this article, the identification of ancestry informative markers (AIMs) within each chromosome has been conducted. The AIM search strategy consists of two main parts: SNP screening and classification model construction. SNPs are screened by a newly proposed round robin symmetrical uncertainty ranking technique and a wrapper embedded with a naive Bayes classifier and a best first search for attribute subset selection. Subsequently, a classification model is built from a naive Bayes classifier. Ten-fold cross-validation is applied during the AIM search. The proposed strategy is implemented and tested on the HapMap Phase II data set, which covers samples from four populations namely the CEU, CHB, JPT and YRI populations [10]. The search for AIMs indicates that the AIM panel which results in complete classification of the HapMap data is not unique. In fact, AIM panels are located on all chromosomes and are made up from lesser number of SNPs than those previously reported [6], [12], [13].

ACKNOWLEDGEMENTS

TP and AA were supported by the Thailand Research Fund (TRF) through the Royal Golden Jubilee Ph.D. Programme (Grant No. PHD/1.E.KN.50/A.1 and PHD/4.I.MU.45/C.1, respectively). TU was supported by the Faculty of Engineering of the King Mongkut's University of Technology North Bangkok. CL was supported by the Mahidol Research Grant, Mahidol University. NC was supported by the Thailand Research Fund.

REFERENCES

- [1] L. Quintana-Murci, O. Semino, H. J. Bandelt, G. Passarino, K. McElreavey, and A. S. Santachiara-Benerecetti, "Genetic evidence of an early exit of *Homo sapiens sapiens* from Africa through Eastern Africa," *Nat. Genet.*, vol. 23, pp. 437–441, 1999.
- [2] M. A. Jobling and C. Tyler-Smith, "The Human Y chromosome: an evolutionary marker comes of age," *Nat. Rev. Genet.*, vol. 4, pp. 598–612, 2003.
- [3] P. Mellars, "Why did modern human populations disperse from Africa ca. 60,000 years ago? A new model," *Proc. Natl. Acad. Sci. USA*, vol. 103, pp. 9381–9386, 2006.
- [4] J. K. Pritchard, M. Stephens, and P. Donnelly, "Inference of population structure using multilocus genotype data," *Genetics*, vol. 155, pp. 945–959, 2000.
- [5] X. Gao and J. Starmer, "Human population structure detection via multilocus genotype clustering," *BMC Genet.*, vol. 8, p. 34, 2007.
- [6] P. Paschou, E. Ziv, E. G. Burchard, S. Choudhry, W. Rodriguez-Cintron, M. W. Mahoney, and P. Drineas, "PCA-correlated SNPs for structure identification in worldwide human populations," *PLoS Genet.*, vol. 3, p. e160, 2007.
- [7] C. Tian, P. K. Gregersen, and M. F. Seldin, "Accounting for ancestry: population substructure and genome-wide association studies," *Hum. Mol. Genet.*, vol. 17, pp. R143–R150, 2008.
- [8] T. M. Baye, H. K. Tiwari, D. B. Allison, and R. C. Go, "Database mining for selection of SNP markers useful in admixture mapping," *BioData Min.*, vol. 2, p. 1, 2009.
- [9] B. Budowle and A. van Daal, "Forensically relevant SNP classes," *BioTechniques*, vol. 44, pp. 603–610, 2008.
- [10] The International HapMap Consortium, "The international HapMap project," *Nature*, vol. 426, pp. 789–796, 2003.
- [11] N. A. Rosenberg, "Algorithms for selecting informative marker panels for population assignment," *J. Comput. Biol.*, vol. 12, pp. 1183–1201, 2005.
- [12] J. Park, S. Hwang, Y. S. Lee, S. C. Kim, and D. Lee, "SNP@Ethnos: a database of ethnically variant single-nucleotide polymorphisms," *Nucleic Acids Res.*, vol. 35, pp. D711–D715, 2007.
- [13] N. Zhou and L. Wang, "Effective selection of informative SNPs and classification on the HapMap genotype data," *BMC Bioinformatics*, vol. 8, p. 484, 2007.
- [14] R. Nunkesser, T. Bernholt, H. Schwender, K. Ickstadt, and I. Wegener, "Detecting high-order interactions of single nucleotide polymorphisms using genetic programming," *Bioinformatics*, vol. 23, pp. 3280–3288, 2007.
- [15] Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, pp. 2507–2517, 2007.
- [16] W. H. Press, B. P. Flannery, S. A. Teukolski, and W. T. Vetterling, *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge: Cambridge University Press, 1988.
- [17] J. R. Quinlan, *C4.5: Programs for Machine Learning*. San Mateo: Morgan Kaufmann, 1993.