Inferring Ethnicity from Mitochondrial DNA Sequence

Chih Lee[1], Ion Mandoiu[1], and Craig E. Nelson[2]

[1] Department of Computer Science and Engineering,
[2] Department of Molecular and Cell Biology, University of Connecticut,
Storrs, CT 06269, USA craig.nelson@uconn.edu

1 Introduction

Ethnicity information can greatly assist forensic investigators, and is also increasingly being used as a predictor of drug effectiveness in the emerging fields of personalized medicine and race-based therapeutics. Self-reported and investigator-assigned ethnicity typically rely on interpreting a complex combination of both genetic and non-genetic markers. However, accurate ethnicity inference can be difficult in contexts with limited access to most informative markers, such as skin and hair samples. The use of genetic information can significantly enhance inference accuracy in these contexts.

Autosomal markers have been shown to provide excellent accuracy for assigning samples to specific clades [9, 11]. Unfortunately, these approaches rely on typing large numbers of autosomal loci that may not survive long periods of degradation. Mitochondrial DNA (mtDNA), however, due to its high-copy number, is recoverable even from minute or highly degraded samples, and, due to its high polymorphism and maternal inheritance, provides excellent information for the inference of ethnic affiliation. Indeed, several studies including [3, 6, 10] have previously shown the feasibility of inferring the probable ethnicity and/or geographic origin from the sequence of the hypervariable region (HVR) of the mitochondrial genome.

In this abstract we study the accuracy of ethnicity inference from mtDNA. As Egeland et al. [6], we consider a supervised learning approach to ethnicity inference. In this setting, mtDNA sequences with annotated ethnicity are used to "train" a classification function that is then used to assign ethnicities to new mtDNA sequences. The main goal of the abstract is to assess the performance of four well-known classification algorithms on a variety of benchmark datasets including realistic levels of missing data and training data bias. The algorithms investigated include support vector machines (SVM) [12], linear discriminant analysis (LDA) [7], quadratic discriminant analysis (QDA) [7], and k-nearest neighbor (kNN). Principal component analysis (PCA) [7], a dimension reduction technique, is used to preprocess the datasets before applying the first three algorithms.

2 Results

To simulate a typical forensic scenario, we extracted from the forensic and published tables in the mtDNA population database [8] samples marked as one of the four coarse ethnic groups – Caucasian, African, Asian and Hispanic. The filtering left 4,426 and 3,976 samples in the forensic and published tables, respectively. These two datasets are referred to as the *forensic* and *published datasets*.

We performed 5-fold cross-validation (CV) analysis using a subset of the forensic dataset called the *trimmed forensic dataset*, which retains only region 16024-16365 of HVR1 from the 4,426 profiles in the forensic dataset. In addition to ethnicity-wise average accuracies, we also used *micro-* and *macro-accuracy* as measures of the overall performance of a classification algorithm. Micro-accuracy is the weighted average of ethnicity-wise accuracy rates by the class sizes, while macro-accuracy is the equally weighted variant. Table 1 summarizes the 5-fold CV accuracy metrics for PCA-QDA, PCA-LDA, 1NN, and PCA-SVM on the trimmed forensic dataset, where PCA-QDA denotes that PCA is applied to the dataset before QDA and so on. PCA-SVM consistently outperforms the other three classification algorithms with respect to all accuracy measures.

Table 1. Comparison of 5-fold CV accuracy measures on the trimmed forensic dataset

	# Samples	Classification Algorithm			
		PCA-QDA	**PCA-LDA**	**1NN**	**PCA-SVM**
Caucasian	1674	83.15	90.2	93.73	94.62
Asian	761	72.93	74.11	83.31	84.76
African	1305	84.6	88.28	86.59	89.81
Hispanic	686	71.57	68.22	72.01	72.59
Micro-Accuracy	4426	80.03	83.46	86.47	88.10
Macro-Accuracy	4426	78.06	80.20	83.91	85.45

Since the performance of different classification algorithms may depend significantly on the typed mtDNA region, we conducted three additional experiments to assess its effect on the classification accuracy of the four compared algorithms. We started from another subset of the forensic dataset called the *full-length forensics dataset*, which consists of 1,904 samples typed for the region of 16024-16569 in HVR1 and 1-576 in HVR2. In the first experiment, we iteratively deleted 10% of the polymorphisms, starting from the HVR2 end non-adjacent to HVR1. Similarly, in the second experiment, we iteratively deleted 10% of the polymorphisms starting from the opposite end. Finally, in the third experiment, we used a sliding window approach to generate 20 different datasets, each of which retains from the full-length forensics profiles 10% of the polymorphisms.

Figures 1 and 2(A) give the 5-fold CV micro-accuracy achieved by PCA-QDA, PCA-LDA, 1NN, and PCA-SVM in these three experiments. Again, PCA-SVM consistently outperforms the other three algorithms. PCA-QDA is typically outperformed by the other methods, except that it outperforms 1NN when the entire HVR is used. 1NN and PCA-LDA have comparable performance, but PCA-LDA performs slightly better than 1NN for near-complete mtDNA profiles. Conversely, 1NN performs better than PCA-LDA for some short typed regions. Indeed, for short windows consisting of only 10% of the polymorphisms in the entire dataset, the performance of 1NN is often as good as that of PCA-SVM, see Fig. 2(A).

Fig. 2(A) further shows that certain regions of HVR1 and HVR2 are more informative than others for the purpose of ethnicity inference. Fig. 2(B) gives the 5-fold CV micro-accuracy for 6 selected windows of 165-271bp spanning the most informative regions of HVR1 and HVR2. Interestingly, when using about 200bp from the information-rich region of HVR1, PCA-SVM yields a microaccuracy of over 80%, very close to the microaccuracy achieved on this set when using the entire HVR region, i.e., HVR1+HVR2.

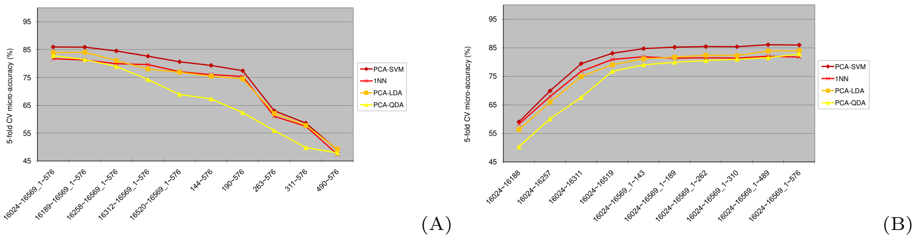

Fig. 1. Comparison of PCA-QDA, PCA-LDA, 1NN, and PCA-SVM 5-fold CV micro-accuracy on regions obtained by iteratively deleting groups of 10% polymorphisms starting from HVR1 towards HVR2 (A), respectively from HVR2 towards HVR1 (B).

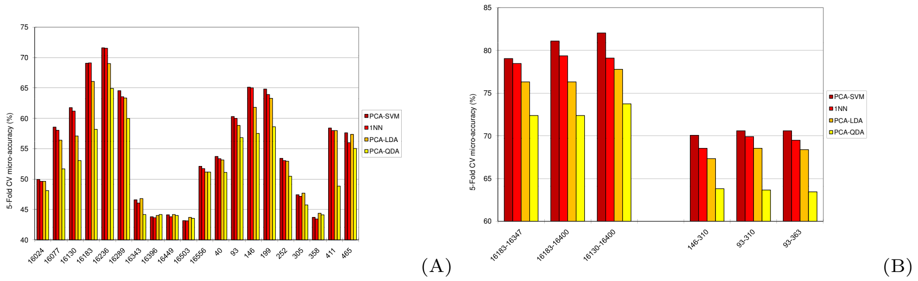

Fig. 2. Comparison of PCA-QDA, PCA-LDA, 1NN, and PCA-SVM 5-fold CV micro-accuracy on sliding windows spanning 10% of the polymorphisms in HVR1+HVR2 (A), and on 6 selected windows of 165-271bp spanning the most informative regions of HVR1 and HVR2 (B).

To obtain a more reliable estimate for the practical accuracy of PCA-SVM, we evaluated its performance using the trimmed forensic dataset as training data and the trimmed published dataset, similarly processed, as test data. Table 2 gives the so called confusion table for this experiment. There is no "Hispanic" row since there are no samples annotated as Hispanic in the test data. However, we do include a “Hispanic” column, because Hispanic samples are present in the training data. PCA-SVM micro-accuracy, as well as ethnicity-wise accuracies for the Caucasian and African ethnic groups are similar to the CV results in Table 2. However, ethnicity-wise accuracy for the Asian group is almost 17% lower than the accuracy achieved in the CV experiment. This is largely explained by large mismatches between Asian profiles used for training and testing in this experiment. The 761 Asian profiles in the Forensic dataset used for training come from only 5 countries: China (356 profiles), Japan (163), Korea (182), Pakistan (8), and Thailand (52), with a strong bias towards East Asia. Not surprisingly, a large percentage of misclassifications errors (90 out of the total of 145) are for profiles collected from two countries (Kazakhstan and Kyrgyzstan) that are not represented in the training dataset. Profiles with unknown country of origin are also poorly classified (10 errors out of 22 samples) suggesting that they may come from regions that are poorly represented in the forensics dataset too.

Table 2. Confusion table of the PCA-SVM test results on the trimmed published dataset

True Ethnicity	# Samples	Predicted Ethnicity			
		Caucasian	Asian	African	Hispanic
Caucasian	1956	**92.59**	5.47	1.53	0.41
Asian	450	25.78	**67.78**	3.11	3.33
African	134	5.22	3.73	**87.31**	3.73
Micro-Accuracy: 87.91%					
Macro-Accuracy: 82.56%					

3 Conclusions

Our experiments show that SVM is the most accurate of compared methods, outperforming both discriminant analysis methods previously employed in [3, 6] as well as a nearest neighbor algorithm similar to that used for haplogroup inference in [2]. In both CV and independent testing, SVM achieves an overall accuracy of 80-90%, matching the accuracy of human experts making ethnicity assignments based on physical measurements of the skull and large bones [4, 5], and coming close to the accuracy achieved by using approximately sixty autosomal loci [1].

References

1. Bamshad, M., Wooding, S., Salisbury, B.A., Stephens, J.C.: Deconstructing the relationship between genetics and race. Nature Reviews Genetics 5(8), 598–609 (2004)

2. Behar, D.M., Rosset, S., Blue-Smith, J., Balanovsky, O., Tzur, S., Comas, D., Mitchell, R.J., Quintana-Murci, L., Tyler-Smith, C., Wells, R.S., Consortium, T.G.: The genographic project public participation mitochondrial dna database. PLoS Genet 3(6), e104 (06 2007)
3. Connor, A., Stoneking, M.: Assessing ethnicity from human mitochondrial dna types determined by hybridization with sequence-specific oligonucleotides. Journal of forensic sciences 39(6), 1360–1371 (1994)
4. Dibennardo, R., Taylor, J.V.: Multiple discriminant function analysis of sex and race in the postcranial skeleton. American Journal of Physical Anthropology 61(3), 305–314 (1983)
5. İşcan, M.Y.: A Topical Guide to the American Journal of Physical Anthropology: Volumes 22-53 (1964-1980). Wiley-Liss (1983)
6. Egeland, T., Bøvelstad, H.M., Storvik, G.O., Salas, A.: Inferring the most likely geographical origin of mtdna sequence profiles. Annals of human genetics 68(5), 461–471 (2004)
7. Hastie, T., Tibshirani, R., Friedman, J.H.: The Elements of Statistical Learning (2nd edition). Springer (2009)
8. Monson, K.L., Miller, K.W.P., Wilson, M.R., DiZinno, J.A., Budowle, B.: The mtdna population database: An integrated software and database resource for forensic comparison. Forensic Science Communications 4(2) (2002)
9. Phillips, C., Salas, A., Sánchez, J., Fondevila, M., Gómez-Tato, A., Álvarez Dios, J., Calaza, M., de Cal, M.C., Ballard, D., Lareu, M., Carracedo, A.: Inferring ancestral origin using a single multiplex assay of ancestry-informative marker snps. Forensic Science International: Genetics 1(3-4), 273 – 280 (2007)
10. Rohl, A., Brinkmann, B., Forster, L., Forster, P.: An annotated mtdna database. International Journal of Legal Medicine 115(1), 29–39 (2001)
11. Shriver, M.D., Smith, M.W., Jin, L., Marcini, A., Akey, J.M., Deka, R., Ferrell, R.E.: Ethnic-affiliation estimation by use of population-specific dna markers. American Journal of Human Genetics 60(4), 957–964 (1997)
12. Vapnik, V.: Statistical Learning Theory. Wiley (1998)