SUMMARY

Statistical methods for estimating genetic distances and phylogenetic trees (dendrograms) from DNA sequence data, restriction data, RAPD data, and microsatellite DNA data were reviewed. A new method for estimating DNA sequence divergence from RAPD data was also presented. It was shown that Nei’s standard genetic distance ($D$) is appropriate for estimating evolutionary time but is less efficient than $D_\alpha$ for constructing phylogenetic trees.

INTRODUCTION

One of the important problems in the conservation of genetic variability in domestic animals is how to choose appropriate breeds for preservation among many available now. One way is to compute the genetic distances between pairs of breeds which are likely to be useful and to choose breeds that cover the widest range of genetic variability. In this case, it is useful to construct phylogenetic trees or dendrograms from distance data and decide which breeds should be kept. Therefore, both the measurement of genetic distances and dendrogram construction become important. In this paper, we discuss the statistical techniques available for these purposes. We will consider the methods for DNA sequence data, restriction data, RAPD data, and microsatellite DNA data, separately.

DNA SEQUENCE DATA

When the genetic distance between different breeds is large or when one wants to know the genetic relationships of different species, DNA sequence data are appropriate. DNA sequences reveal the ultimate source of genetic variability, and thus they are ideal data for measuring genetic distances. In practice, however, it is time-consuming and expensive to gather extensive sequence data. Therefore, we are usually forced to examine a relatively short stretch of DNA, which is highly variable. One such piece of DNA is the D-loop region of mitochondrial DNA (mtDNA). This region of mtDNA is now being used extensively for studying the genetic differentiation of human populations (e.g., Stoneking 1993).

Measures of genetic distance

The simplest measure of genetic distance for DNA sequence data is the proportion of nucleotide differences, i.e., $p$-distance. This $p$-distance is defined by

$$p = \frac{n_d}{n},$$

where $n$ is the number of nucleotides examined and $n_d$ is the number of nucleotide differences.

This distance measure, however, does not take into account multiple nucleotide substitutions at the same site. To correct for these multiple substitutions, we have to use a mathematical model of nucleotide substitution. The simplest model available for this purpose is Jukes-Cantor’s (1969), where the rate of substitution is assumed to be the same for all pairs of the four nucleotides A, T, C, and G. The number of nucleotide substitutions per site ($d$) under this model can be estimated by
\[ \hat{d} = -\frac{3}{4} \ln \left(1 - \frac{4}{3}p\right), \]  

where \( p \) is the proportion of different nucleotides. In practice, Jukes and Cantor's model is not necessarily realistic, but as long as the \( d \) is smaller than 0.05, \( \hat{d} \) is known to be a good estimate.

In actual sequence data, the rate of transitional nucleotide substitution (A \( \leftrightarrow \) G and T \( \leftrightarrow \) C) is often higher than that of transversional substitution (all other substitutions). This is particularly so for animal mtDNA. In this case, Kimura's (1980) 2-parameter method gives a better estimate of \( d \). That is,

\[ \hat{d} = -\frac{1}{2} \ln \left(1 - 2P - Q\right) - \frac{1}{4} \ln \left(1 - 2Q\right), \]

where \( P \) and \( Q \) are the proportions of transitional and transversional nucleotide differences.

Phylogenetic trees

There are many statistical methods for constructing phylogenetic trees or dendrograms from DNA sequence data. They can be classified into distance methods and discrete-character methods. In distance methods, a pair-wise distance is computed for all sequences, and a phylogenetic tree is constructed from distance data by certain principles and algorithms. Currently popular distance methods are the unweighted pair-group method with arithmetic means (UPGMA; Sneath and Sokal 1973) and the neighbor-joining (NJ) method (Saitou and Nei 1987). The NJ method is a simplified version of the minimum evolution method later developed by Saitou and Imanishi (1989) and Rzhetsky and Nei (1992). UPGMA depends on the assumption that the rate of evolution is the same for all evolutionary lineages, but the NJ method requires no such assumption. In discrete-character methods, a tree is constructed by considering the evolutionary relationships of DNA sequences at each nucleotide site. Two major methods belonging to this category are the maximum parsimony (MP) method (Eck and Dayhoff 1966) and the maximum likelihood (ML) method (Felsenstein 1981).

According to computer simulations, the MP method is known to be efficient in obtaining the correct topology when the number of nucleotide substitutions per site is small (<0.05) and the number of nucleotides examined is large (Nei 1991), but in general it is less efficient than the NJ or the ML method (Saitou and Imanishi 1989; Tateno et al. 1994). The NJ and ML methods seem to be nearly equally efficient and better than most other tree-building methods. However, the NJ method requires much less computer time than the ML method.

RESTRICTION DATA

The extent of DNA polymorphism or sequence divergence can also be studied by using restriction enzymes (Nei 1987). If a circular DNA such as animal mtDNA has \( m \) recognition (restriction) sites for a restriction enzyme, it is fragmented into \( m \) segments after digestion by this enzyme. A restriction enzyme recognizes a specific sequence of nucleotide pairs, generally four or six pairs in length, and cleaves it. The number and location of restriction sites vary with nucleotide sequence. The higher the similarity of the two sequences, the closer the cleavage pattern. Therefore, it is possible to estimate the
number of nucleotide substitutions between two homologous DNAs by comparing the location of restriction sites.

However, the statistical methods for estimating the number of nucleotide substitutions have already been discussed elsewhere. Therefore, we shall not repeat them here. A reader who is interested in these methods may refer to Nei (1987) and Nei and Miller (1990). The only comment we want to make here is that this method is efficiently used only for the case where $d$ is equal to or smaller than 0.1.

**RAPD DATA**

A few years ago, random amplification of polymorphic DNA (RAPD) by the polymerase chain reaction (Welsh and McClelland 1990; Williams et al. 1990) was hailed as a future technique of studying DNA polymorphism. At the present time, this enthusiasm has subsided to some extent, because the reproducibility of the results is not always high. However, if we use only RAPD results that are reproducible, they seem to be useful for obtaining a rough estimate of genetic divergence. One advantage of this method is that a large set of data can be obtained rapidly (Clark and Lanigan 1993; Tibayrenc et al. 1993).

Recently Clark and Lanigan (1993) developed a statistical method for estimating the number of nucleotide substitutions under the assumption that the oligonucleotide primers (usually about 10 nucleotides long) used hybridize to all complementary sites in the DNA sequence and that one or more base mismatches preclude hybridization. Although this assumption may not hold in practice, there seems to be some way to eliminate data which do not satisfy the assumption (Tibayrenc et al. 1993).

Clark and Lanigan’s method is to estimate the proportion ($F$) of DNA fragments that are shared by two populations and to use Nei and Li’s (1979) formula to estimate the average number of nucleotide substitutions per site ($d$). Nei and Li have shown that if $P$ is the probability of no mutation occurring at a primer site with $r$ nucleotides for a period of $t$ years, it can be expressed as $P = \exp(-rt\lambda)$, where $\lambda$ is the rate of nucleotide substitution per site per year. They then showed that $P$ can be estimated by solving the following equation.

$$P = [F(3 - 2P)]^{1/4}.$$  

This equation can be solved by an iterative computation of $P = [F(3 - 2P_1)]^{1/4}$, where $P_1$ is a trial value. Nei (1987) has suggested that $P_1 = F^{1/4}$ be used as the initial trial value. Once $P$ is obtained, the estimate of $d$ is given by

$$\hat{d} = - (2/r) \ln P.$$  

Clark and Lanigan presented a formula for estimating $F$ for RAPD data. However, their formula is not very accurate and is expected to have a large variance, though their computer simulation suggests that it gives a quite good estimate of $d$ when sample size is large. An accurate formula is given by

$$\hat{F} = \frac{\sum p_{xi} p_{yi}}{\left(\sum p_{xi}^2 \cdot \sum p_{yi}^2\right)^{1/2}}.$$  

where $p_{Xi}$ and $p_{Yi}$ are the frequency of the $i$-th DNA fragment (electrophoretic band) in population $X$ and $Y$, respectively. The rationale behind this equation is as follows. In RAPD experiments, only DNA fragments or electrophoretic bands of 500 - 3,500 nucleotides are detected (Clark and Lanigan 1993), and each band represents one segment of the genome under investigation, and we are interested only in the bands which are polymorphic. In other words, there must be some individuals having a band and some individuals lacking the band in a population, and the frequency of the genome having the $i$-th band is
\( p_{X_i} \) in population \( X \) and \( p_{Y_i} \) in population \( Y \) as mentioned above. Therefore, the probability that the two populations share the band is \( p_{X_i} p_{Y_i} \). If there are \( b \) bands detected in an experiment, the sum of \( p_{X_i} p_{Y_i} \) is given by \( \sum_{i=1}^{b} p_{X_i} p_{Y_i} \). However, we must normalize this quantity by the probability of two randomly chosen genome from the same population having the same band. This probability is given by \( \sum_{i=1}^{b} p_{X_i}^2 \), in population \( X \) and \( \sum_{i=1}^{b} p_{Y_i}^2 \), in population \( Y \). Therefore, we obtain equation (6).

In haploid populations or homozygous diploid populations, \( F \) can easily be computed. In diploid populations, however, an individual having a band may be homozygous or heterozygous. Therefore, \( p_{X_i} \) must be estimated. If we assume Hardy-Weinberg equilibrium, \( p_{X_i} \) can be estimated by \( p_{X_i} = 1 - Q_i^b \), where \( Q_i \) is the frequency of individuals lacking the \( i \)-th band. Clark and Lanigan (1993) have shown that this method gives sufficiently good estimates of \( d \).

**MICROSATELLITE DNA DATA**

During the last two decades, protein polymorphism data have been used extensively for studying genetic distances, but they are not always useful for studying closely related populations because the extent of polymorphism is often quite low. Recent discovery of microsatellite DNA has changed this situation. Microsatellite DNA has high mutation rates (Weber and Wong 1993), and the extent of polymorphism is so high that the average heterozygosity ranges from 30% to 80%. Edwards et al. (1992) and Valdes et al. (1993) have shown that the pattern of mutation follows a kind of stepwise mutation model, which is very close to the infinite-allele model of mutation when a relatively short evolutionary time is considered. Therefore, these data can be analyzed by the same statistical methods as those used for protein polymorphism data.

There are many different measures of genetic distances for allele frequency data (Nei 1987), and the merits and demerits of these measures have been disputed. Genetic distances are used either for estimating evolutionary times or for constructing phylogenies (dendrograms), and a distance measure that is appropriate for one purpose is not necessarily good for the other (Nei et al. 1983). We have therefore examined the linear relationships with evolutionary time and the probability of obtaining the correct tree by using computer simulation. We examined nine different measures of genetic distance presented in Nei (1987). They were \( D_s \) (Rogers 1972), \( C_p \) (Prevosti et al. 1975), \( D_a \) (Nei and Roychoudhury 1972), \( D \) (Nei 1972), \( D_L = - \ln (1 - \phi^* \) (Latter 1972), \( D_c \) (Cavalli-Sforza and Edwards 1967), \( f_6 \) (Cavalli-Sforza 1969), \( D_\alpha \) (Nei et al. 1983), and \( Q^2 \) (Bhattacharyya 1946; Nei 1987). In the present simulation, we sampled all individuals in the populations, so that Reynolds et al.'s (1983) \( D_w \) is identical with \( D_L \).

The procedure of our computer simulation was similar to that of Nei et al. (1983). We assumed that a population splits into two identical populations at an evolutionary time and that the two populations evolve independently by mutation and genetic drift. This process was continued until eight populations were generated following the model tree in Fig 1. We also assumed that the populations are in mutation-drift equilibrium throughout the evolutionary process. We considered two levels of average heterozygosity, i.e., \( H = 0.16 \) and \( H = 0.5 \). The heterozygosity value of \( H = 0.16 \) is appropriate for electrophoretic alleles and is the same as that used by Nei et al.'s (1983) simulation. \( H = 0.5 \) seems to be appropriate for microsatellite DNA data. To facilitate the computer simulation, we used the effective population size \( (N) \) equal to 50 and the mutation rate per locus per generation \( (v) \) equal to 0.001 for the case of \( H = 0.16 \) and \( N = 250 \) and \( v = 0.001 \) for the case of \( H = 0.5 \).

Following Nei et al. (1983), we generated allele frequency data for 100 loci for each of the eight populations and computed the nine different genetic distances for the first 10, 20, 30, ..., and 100 loci separately. Using these distance data, we then constructed a phylogenetic tree by UPGMA or the NJ method and examined whether the topology of the reconstructed tree agreed with that of the model tree or not. This was repeated 100 times, and the percentage of replications in which the correct topology was obtained was computed for each set of loci and for each distance measure used.

Relationships of genetic distances with evolutionary time
Fig. 1. Model tree for computer simulation. The branch lengths are measured in terms of the expected number of gene substitutions ($v_t$).

**Table 1. Percentage of replications in which the correct topology was obtained**

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$H = 0.16$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>5 (5)</td>
<td>6 (5)</td>
<td>2 (7)</td>
<td>4 (5)</td>
<td>2 (8)</td>
<td>7 (9)</td>
<td>1 (3)</td>
<td>5 (8)</td>
</tr>
<tr>
<td>20</td>
<td>10 (18)</td>
<td>17 (24)</td>
<td>15 (24)</td>
<td>16 (9)</td>
<td>4 (22)</td>
<td>14 (20)</td>
<td>7 (4)</td>
<td>19 (26)</td>
</tr>
<tr>
<td>30</td>
<td>20 (28)</td>
<td>23 (32)</td>
<td>21 (31)</td>
<td>23 (18)</td>
<td>15 (34)</td>
<td>22 (33)</td>
<td>10 (13)</td>
<td>27 (44)</td>
</tr>
<tr>
<td>40</td>
<td>32 (40)</td>
<td>30 (46)</td>
<td>39 (44)</td>
<td>35 (32)</td>
<td>21 (48)</td>
<td>35 (44)</td>
<td>20 (17)</td>
<td>40 (53)</td>
</tr>
<tr>
<td>50</td>
<td>39 (47)</td>
<td>39 (56)</td>
<td>44 (55)</td>
<td>38 (39)</td>
<td>22 (56)</td>
<td>39 (53)</td>
<td>29 (23)</td>
<td>48 (63)</td>
</tr>
<tr>
<td>60</td>
<td>50 (53)</td>
<td>53 (63)</td>
<td>55 (57)</td>
<td>50 (42)</td>
<td>31 (61)</td>
<td>52 (58)</td>
<td>34 (28)</td>
<td>53 (67)</td>
</tr>
<tr>
<td>70</td>
<td>60 (62)</td>
<td>61 (69)</td>
<td>62 (68)</td>
<td>63 (58)</td>
<td>38 (70)</td>
<td>63 (69)</td>
<td>40 (37)</td>
<td>59 (77)</td>
</tr>
<tr>
<td>80</td>
<td>63 (65)</td>
<td>64 (72)</td>
<td>65 (73)</td>
<td>67 (61)</td>
<td>37 (77)</td>
<td>65 (74)</td>
<td>42 (37)</td>
<td>71 (84)</td>
</tr>
<tr>
<td>90</td>
<td>68 (72)</td>
<td>70 (80)</td>
<td>70 (79)</td>
<td>68 (62)</td>
<td>45 (82)</td>
<td>71 (79)</td>
<td>40 (44)</td>
<td>76 (88)</td>
</tr>
<tr>
<td>100</td>
<td>73 (75)</td>
<td>76 (79)</td>
<td>77 (81)</td>
<td>78 (70)</td>
<td>47 (84)</td>
<td>76 (81)</td>
<td>43 (50)</td>
<td>84 (86)</td>
</tr>
<tr>
<td>$H = 0.5$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>8 (7)</td>
<td>14 (15)</td>
<td>9 (9)</td>
<td>8 (4)</td>
<td>10 (11)</td>
<td>20 (19)</td>
<td>12 (12)</td>
<td>20 (24)</td>
</tr>
<tr>
<td>20</td>
<td>20 (20)</td>
<td>30 (29)</td>
<td>24 (21)</td>
<td>19 (13)</td>
<td>25 (21)</td>
<td>38 (38)</td>
<td>22 (19)</td>
<td>45 (43)</td>
</tr>
<tr>
<td>30</td>
<td>30 (35)</td>
<td>54 (50)</td>
<td>49 (37)</td>
<td>47 (23)</td>
<td>35 (41)</td>
<td>63 (58)</td>
<td>45 (46)</td>
<td>64 (59)</td>
</tr>
<tr>
<td>40</td>
<td>36 (50)</td>
<td>64 (56)</td>
<td>59 (54)</td>
<td>50 (33)</td>
<td>49 (59)</td>
<td>70 (71)</td>
<td>54 (55)</td>
<td>75 (74)</td>
</tr>
<tr>
<td>50</td>
<td>62 (55)</td>
<td>66 (67)</td>
<td>65 (55)</td>
<td>61 (36)</td>
<td>55 (60)</td>
<td>72 (71)</td>
<td>63 (61)</td>
<td>77 (75)</td>
</tr>
<tr>
<td>60</td>
<td>65 (64)</td>
<td>79 (75)</td>
<td>73 (67)</td>
<td>66 (48)</td>
<td>63 (69)</td>
<td>83 (82)</td>
<td>68 (66)</td>
<td>85 (83)</td>
</tr>
<tr>
<td>70</td>
<td>78 (67)</td>
<td>81 (76)</td>
<td>81 (70)</td>
<td>72 (55)</td>
<td>73 (72)</td>
<td>88 (81)</td>
<td>74 (74)</td>
<td>93 (86)</td>
</tr>
<tr>
<td>80</td>
<td>81 (74)</td>
<td>87 (81)</td>
<td>85 (74)</td>
<td>82 (60)</td>
<td>77 (74)</td>
<td>89 (88)</td>
<td>74 (75)</td>
<td>91 (89)</td>
</tr>
<tr>
<td>90</td>
<td>86 (73)</td>
<td>87 (86)</td>
<td>89 (76)</td>
<td>81 (61)</td>
<td>80 (78)</td>
<td>91 (92)</td>
<td>81 (77)</td>
<td>95 (94)</td>
</tr>
<tr>
<td>100</td>
<td>86 (77)</td>
<td>91 (89)</td>
<td>91 (82)</td>
<td>89 (71)</td>
<td>88 (82)</td>
<td>91 (93)</td>
<td>82 (79)</td>
<td>96 (94)</td>
</tr>
</tbody>
</table>

Note. - UP: UPGMA. NJ: Neighbor-joining method.
Fig. 2. Relationship of various genetic distance measures with evolutionary time
We first examined the relationship of each distance measure with evolutionary time ($vt$, where $t$ is the number of generations) using the distance data for 100 loci. The results obtained are presented for six selected distances (Fig.2). (All other distances showed a clear nonlinear relationship with evolutionary time.) As is theoretically known (Nei 1972), $D$ increases linearly irrespective of the average heterozygosity. All other distances show a nonlinear relationship, but the deviation from linearity varies with distance measure. This deviation is quite large in $D_c$ and $C_P$ irrespective of the $H$ value. The deviation in $D_2$ is somewhat large for $H = 0.16$ but becomes smaller for $H = 0.5$. The curves for $D_A$ and $\theta^2$ are nonlinear but close to linearity for both $H = 0.16$ and $H = 0.5$. These results indicate that $D$ is best for estimating evolutionary time, but $D_A$ and $\theta^2$ can also be used for this purpose if the evolutionary time considered is short.

Probability of obtaining the correct topology

Table 1 shows the percentages ($P_c$) of replications in which the correct topology was obtained for each distance measure. The $P_c$ values for $\theta^2$ were virtually identical with those of $D_A$, so that they are not presented here. When UPGMA is used, the $P_c$ value for $H = 0.16$ is more or less the same for $D_p$, $C_p$, $D_m$, $D$, and $D_c$ but is lower for $D_2$ and $f_6$. By contrast, $D_A$ shows the highest $P_c$ value for most sets of loci. When $H = 0.5$ is assumed, however, the $P_c$ for $D_c$ increases substantially and is slightly higher than that of $D_p$, $C_p$, $D_m$, and $D$. The $P_c$ for $f_6$ also increases considerably. However, the best distance measure for obtaining a high $P_c$ value is $D_A$.

When the NJ method is used, the $P_c$ for $H = 0.16$ is almost always higher than that for the case of UPGMA. One exception is $D$, for which UPGMA often shows a higher $P_c$ value than the NJ method. For $H = 0.5$, however, NJ gives lower $P_c$ values than UPGMA for almost all distance measures. [When some populations are subject to bottlenecks, NJ is generally better than UPGMA even for $H = 0.5$ (M. Nei and N. Takezaki, unpublished).] For the case of $H = 0.16$, $f_6$ again shows a poor performance, but the $P_c$ for this distance increases substantially when $H = 0.5$. Curiously, $D$ shows the poorest performance when $H = 0.5$. The best distance measure for obtaining the correct tree is again $D_A$ in both cases of $H = 0.16$ and $H = 0.5$.

The above results are more or less the same as those obtained by Nei et al. (1983), though these authors examined only $D$, $D_p$, $D_m$, $f_6$, and $D_A$ using UPGMA. Therefore, it now seems established that the genetic distance that is appropriate for estimating evolutionary time is not necessarily the best for obtaining the correct topology. If one is interested in estimating the correct topology, the best distance measure seems to be either $D_A$ or $\theta^2$. In practice, the mathematical expression for $D_A$ is simpler than that for $\theta^2$, so that $D_A$ is preferable for practical use.

DISCUSSION

In this paper we considered only theoretical aspects of estimating genetic distances and phylogenetic trees. We have shown that the distance that is appropriate for estimating evolutionary time or accumulated amounts of gene substitutions is not necessarily good for phylogenetic inference. What should we do then? One solution for this problem would be to separate topology estimation and branch length estimation. Thus, we can estimate the topology of a tree by using $D_A$ but estimate branch lengths by using $D$ for a given topology. Theoretically this can be done if we use the least-squares method for estimating branch lengths for a given topology (Rzhetsky and Nei 1992). In this particular case, however, we need a simple procedure of estimating the variance-covariance matrix of $D$ distances. We are now studying this problem.

Once a reasonable phylogenetic tree (or dendrogram) is produced, we can decide which animal breeds to be maintained under the condition that the largest amount of genetic variability is to be retained. In practice, however, this procedure alone would not be sufficient. If a certain breed has rare morphological or physiological characters, it should also be preserved. Therefore, the conservation of genetic variability of domestic animals should be conducted by considering various aspects of biological
factors.

Computer programs

The computer programs MEGA and DISPAN (T. Ota, 1993) are available upon request. The former program is for estimating genetic distances and phylogenetic trees from DNA or protein sequences data, whereas the latter is for computing $D$ and $D_A$ and constructing UPGMA and NJ trees from allele frequency data.

REFERENCES

TIBAYRENC, M., NEUBAUER, K., BARNABE, C., GUERRINI, F. SKARECKY, D. and AYALA, F.J.