In Boyce A. J. and Mascie-Taylor C.G.N. (eds.)
Molecular Biology and Human Diversity, pp. 265-282
Cambridge University Press, Cambridge (1996)

18 *Contrasting gene trees and population trees of the evolution of modern humans*

N. SAITOU

A gene tree is an essential descriptor of any evolutionary process, for the semi-conservative replication of the DNA double helix automatically produces a bifurcating gene tree. It should be emphasized that the genealogical relationship of genes is independent of the mutation process, especially when neutral evolution (Kimura, 1983) is considered. The former is a direct product of DNA replication, while the latter may or may not happen within a certain time period and DNA region. Therefore, even if many nucleotide sequences happened to be identical, there must be a genealogical relationship for those sequences. However, it is impossible to reconstruct the genealogical relationship without mutational events. In this respect, extraction of mutations from genes and their products is critical for reconstructing phylogenetic trees.

We can, therefore, best estimate a gene tree according to the mutation events realized on its expected gene tree (see Fig. 18.1(*a*)). We call this ideal reconstruction of the gene tree the realized gene tree (see Fig. 18.1(*b*)), while the reconstructed one from observed data is called the 'estimated' gene tree (Saitou, 1995*b*). Branch lengths of realized and estimated genes tree are proportional to mutational events. These mutational events are not necessarily proportional to physical time. Because of limitations in information, estimated gene trees are often unrooted. By definition, expected gene trees are strictly bifurcating, while realized and estimated gene trees may be multifurcating. This is because of the possibility of no mutation at a certain interior branch, such as branch X of Fig. 18.1(*a*).

Ideally, branch lengths of a phylogenetic tree are proportional to physical time. We call this type of tree the 'expected tree'. It is a rooted tree. Both species/population trees and gene trees have their expected trees, but their properties are somewhat different from each other. An expected gene tree directly reflects the history of DNA replications. In contrast, a

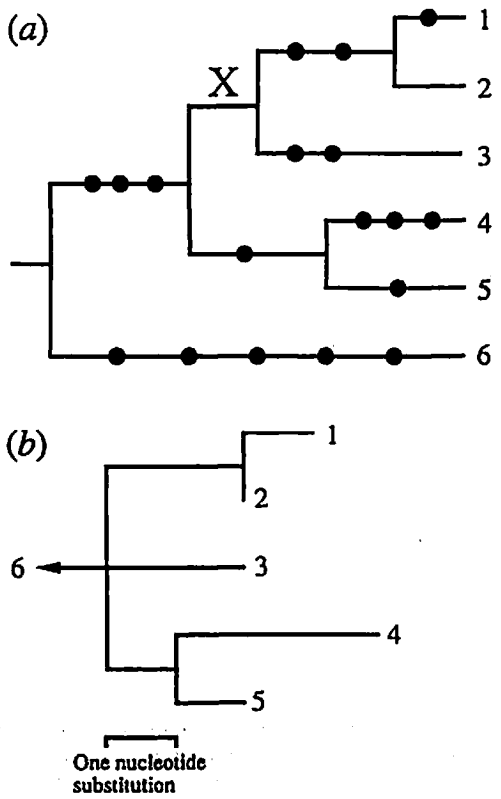

Fig. 18.1. Expected (*a*) and realized (*b*) gene tree (From Saitou (1995*b*)). Full circles on the expected gene tree denote nucleotide substitutions. Because no substitution occurred at branch X of the expected gene tree (*a*), the corresponding branch does not exist in the realized gene tree (*b*).

species/population tree is only a simplified view of a nexus of gene trees. Therefore, the speciation time (or time of population diversification) is not always clear, in contrast to the clear DNA replication event. A species/population tree reconstructed from observed data is called an 'estimated' species/population tree, while there is no realized species/population tree.

There are several other important differences between gene trees and species/population trees. Even when orthologous genes are used, a gene tree may be different from the corresponding species/population tree. This difference arises from the existence of a gene genealogy in ancestral species/population. A simple example is illustrated in Fig. 18.2. A gene sampled from species A has its direct ancestor at the speciation time T_1 generations ago, and so does a gene sampled from species B. Thus the divergence time between the two genes sampled from the different species always overestimates that of the species. The amount of overestimation corresponds to the coalescence time in the ancestral species, and its

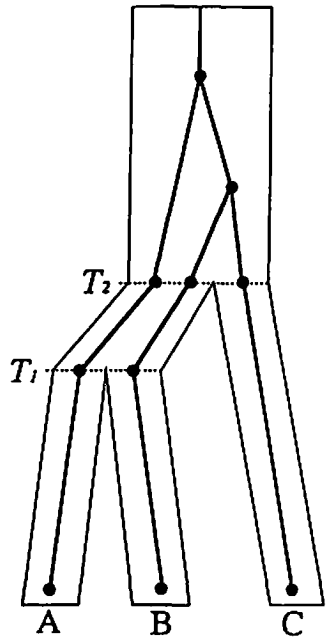

Fig. 18.2. Difference between a gene genealogy and species tree (From Saitou, 1996). Full circles and thick lines denote the genealogical relationship while thin lines (outlining the gene tree) denote the species tree. A, B, and C denote three genes each sampled from extant species, while X and Y denote ancestral genes. T_1 and T_2 denote the two speciation times.

expectation is $2N$ for neutrally evolving nuclear genes of a diploid organism, where N is the population size of the ancestral species. Therefore, if the two speciation events (T_1 and T_2) are close enough, the topological relationships of the gene tree may become different from those of the species tree, as shown in Fig. 18.2. Although species A and B are more closely related than to C, genes sampled from species B and C happen to be more closely related to each other than to that sampled from species A. The probability (P_{error}) of obtaining an erroneous tree topology is given by $P_{error} = (2/3)e^{-T/2N}$, where $T = T_2 - T_1$ generations (Nei, 1987). For example, P_{error} is 0.404 when $T = 50\,000$ and $N = 50\,000$. Therefore, a species tree estimated from a single locus may not be correct even if the gene tree has been correctly estimated; we should use more than one locus. Saitou and Nei (1986) computed the probabilities of obtaining the correct species tree from a number of gene trees for the case of a three species tree. They considered a trinomial distribution, and the topology supported by the largest number of loci was regarded as the correct one. Under this condition, we need only one locus when $T/2N$ is 4, but 7 loci when $T/2N$ is 1, if we want the probability to be larger than 0.95.

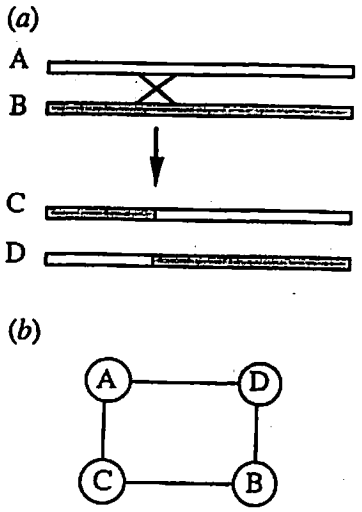

Fig. 18.3. (*a*) A recombination event creates new alleles C and D from existing alleles A and B. (*b*) A network of four alleles caused by the recombination described in (*a*).

When gene conversion and/or recombination has occurred within the gene region under consideration, a tree structure may no longer exist. Fig. 18.3 shows this situation schematically. When a recombination occurs between alleles A and B that diverged some time ago, new recombinant alleles C and D are produced (Fig. 18.3(*a*)). If we consider the relationship among those four alleles, the resultant graph is not a tree but a network (Fig. 18.3(*b*)). The distance (measured in terms of nucleotide difference) between alleles A and C is the same as that between B and D, and it is smaller than that between A and D (or between B and C) if we consider the location of the recombination event. It should be noted that the mutations are assumed to have accumulated uniformly over the sequence. This example clearly shows the limitation of a tree representation in some cases. Bandelt (1994) recently proposed a method for constructing such networks from sequence data.

A gene tree for HTLV-I DNA

Human T-lymphotropic virus type I (HTLV-I) has been found in Japan, Africa, and the Caribbean Islands, but it has also been found in Melanesia (see Yanagihara & Garruto, 1992 for a review and Yanagihara *et al.*, 1995). Nerurkar *et al.* (1993) sequenced parts of the HTLV-I genome found in Melanesians (Papua New Guineans and Solomon Islanders), and Song *et al.* (1994) determined several simian T-lymphotropic virus type I (STLV-I)

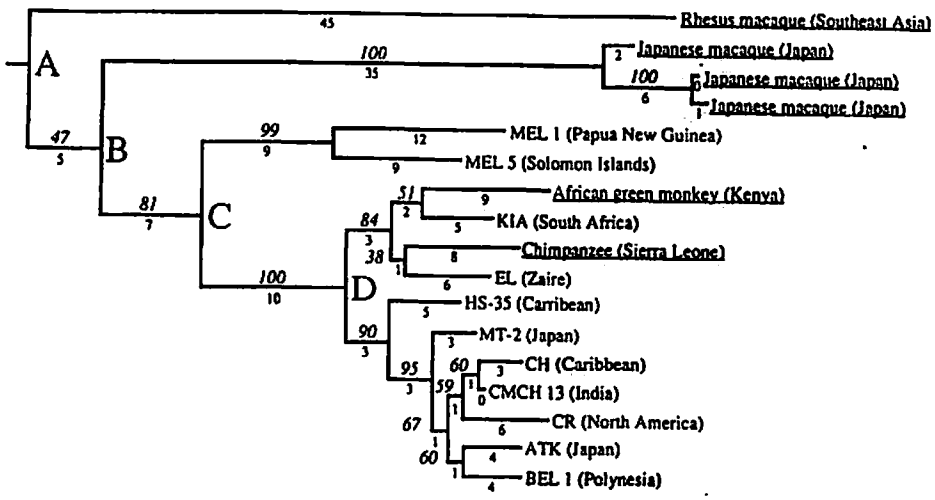

Fig. 18.4. A neighbour-joining tree of HTLV-I and STLV-I sequences (modified from Song *et al.*, 1994). The tree is rooted by including an HTLV-II sequence. STLV-I sequences are underlined. Numbers below branches are estimated numbers of nucleotide substitutions at corresponding branches, and those above internal branches are bootstrap probabilities (%).

sequences. A phylogenetic tree of HTLV-I and STLV-I sequences is shown in Fig. 18.4. HTLV-II which is remotely related to HTLV-I and STLV-I was used as an outgroup to locate the root of the tree. Because we are interested in reconstructing the realized gene tree, all the estimated branch lengths (integers below branches) are numbers of nucleotide substitutions that have occurred in the compared DNA region.

First of all, it is evident that the resultant gene tree does not correspond to the phylogenetic tree of species involved in the comparison (human, chimpanzee, African green monkey, Japanese macaque, and rhesus macaque). Because the two macaque species and African green monkey are Old World monkeys, they should be monophyletic in the real species tree. Accordingly, the branching point (node) A of Fig. 18.4 is unlikely to correspond to the separation time of rhesus and Japanese macaques, nor node B to the separation time of Japanese macaques and humans. Alternatively, it may be more reasonable to assume that node C corresponds to the time of the first human migration into Melanesia (*ca.* 50 000 BP) from Sunda land. If so, node D, the coalescent point for the so-called cosmopolitan HTLV-I strains, corresponds to about 25 000 BP under the assumption of a rough constancy of the evolutionary rate. Nodes A and B are also roughly dated as *ca.* 90 000 BP and 70 000 BP, respectively. Both human and macaques cohabited in the Asian Continent around that time, and it is conceivable that an interspecific transmission of the virus occurred. Similar interspecific transmissions evidently occurred much more recently

in Africa, for African green monkey and chimpanzee STLV-I strains clustered with some human sequences from Africa (see Fig. 18.4).

In spite of some initial enthusiasm, there are doubts as to the utility of the phylogenetic tree of the HTLV-I virus for elucidating modern human evolution. As shown above, the interspecific transmission of this virus between human and other primates seems to occur rather frequently. If so, horizontal transfer of the virus among humans may easily occur. Thus the HTLV-I/STLV-I gene trees should not be used without due consideration.

Gene trees and population trees for mitochondrial DNA

Genetic polymorphism of human mitochondrial DNA (mtDNA) has been extensively studied by using both restriction enzymes and direct sequencing. We studied both gene trees and population trees based on mtDNA polymorphisms detected by restriction enzymes (Harihara & Saitou, 1989; Saitou & Harihara, 1996). A summary of the results is given in this section.

Published mtDNA data were collected for a total of 885 individuals from 15 human populations (see Harihara & Saitou, 1989 for details). The restriction enzymes used were *Ava* II, *Bam* HI, *Hpa* I, and *Msp* I. Each enzyme produces various patterns of restriction fragments, and the sets of restriction sites deduced from such patterns are called mtDNA morphs. Fig. 18.5 shows the relationship of 26 mtDNA morphs found by using *Ava* II. It is clear that morph 1 is at the centre of radiation, followed by morph 5. When there is more than one possibility of connecting different morphs with more than one restriction site difference (e.g. morphs 1 and 20), a loop is created. This is reminiscent of the phylogenetic network.

A combination of each mtDNA morph for different restriction enzymes is called a mtDNA type. A total of 57 mtDNA types (or haplotypes) were found by this procedure. Although we observed a network structure for mtDNA morphs, the real evolutionary history of mtDNA molecules must be a phylogenetic tree, for mtDNA is considered to undergo no recombination. We therefore produced the mtDNA gene tree (Fig. 18.6) using the maximum parsimony method (Fitch, 1977). It should be mentioned that the tree shown in Fig. 18.6 is only one of many equally parsimonious trees.

MtDNA type 1 was the most frequent (670 individuals out of 885 had this type), and was found in all 15 populations. This cosmopolitan mtDNA type is shown as the large ellipse in Fig. 18.6, and is at the centre of radiation. Most of the mtDNA types found in African populations (Bantu and Bushman) form a monophyletic cluster at the left side of branch α, although some mtDNA types found in Arabian (P) and Roman (R) populations are also included in this 'African' cluster. This clear distinction

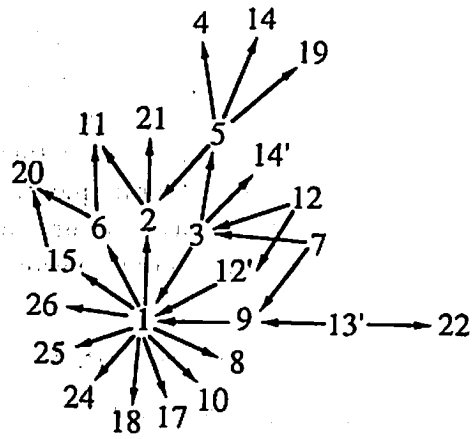

Fig. 18.5. A phylogenetic network for 26 mitochondrial morphs produced by using restriction enzyme *Ava* II (modified from Fig. 18.1a of Harihara & Saitou, 1989). Arrows indicate the direction of restriction site loss.

of mtDNA types found in African populations is a basis of the 'Out-of-Africa' hypothesis championed by Cann, Stoneking & Wilson (1987) using their restriction site data, and later extended by Vigilant *et al.* (1991) using nucleotide sequence data. However, re-analysis of these data (Hedges *et al.*, 1991; Madisson, Ruvolo & Swofford, 1992) showed that there is still uncertainty for the support of this hypothesis.

Because the gene tree of Fig. 18.6 does not have a root, we assigned a root to mtDNA type I (cosmopolitan type) for the following four reasons (Saitou & Harihara, 1996). (1) Under neutral evolution, the most frequent allele (mtDNA type 1) is likely to be the oldest, with a probability equal to its frequency (Watterson & Guess, 1977). In this case, the frequency of mtDNA type 1 is 670/885 = 0.76. (2) All 15 populations had mtDNA type 1. (3) mtDNA type 1 is the centre of radiation of the remaining mtDNA types; there were 23 branches connected to mtDNA type 1. (4) If we use the midpoint rooting method assuming a rough constancy of evolutionary rate (Farris, 1972), mtDNA type 1 becomes the root.

A rooted tree of 56 mtDNA types was thus obtained. We then estimated the evolutionary rate of mtDNA as follows: the average number of restriction site differences between the ancestral mtDNA type (type 1) and present-day individuals was computed to be 0.514. Because the mtDNA type 1 was a combination of *Ava* II morph 1 (8 restriction sites), *Bam* HI morph 1 (1 restriction site), *Hpa* I morph 2 (3 restriction sites), and *Msp* I morph 1 (23 restriction sites), the total number of nucleotides assayed by these four restriction enzymes is $8 \times 5 + (1 + 3) \times 6 + 23 \times 4 = 156$. Thus the average number of nucleotide differences per nucleotide site between the ancestral mtDNA type and present-day individuals was

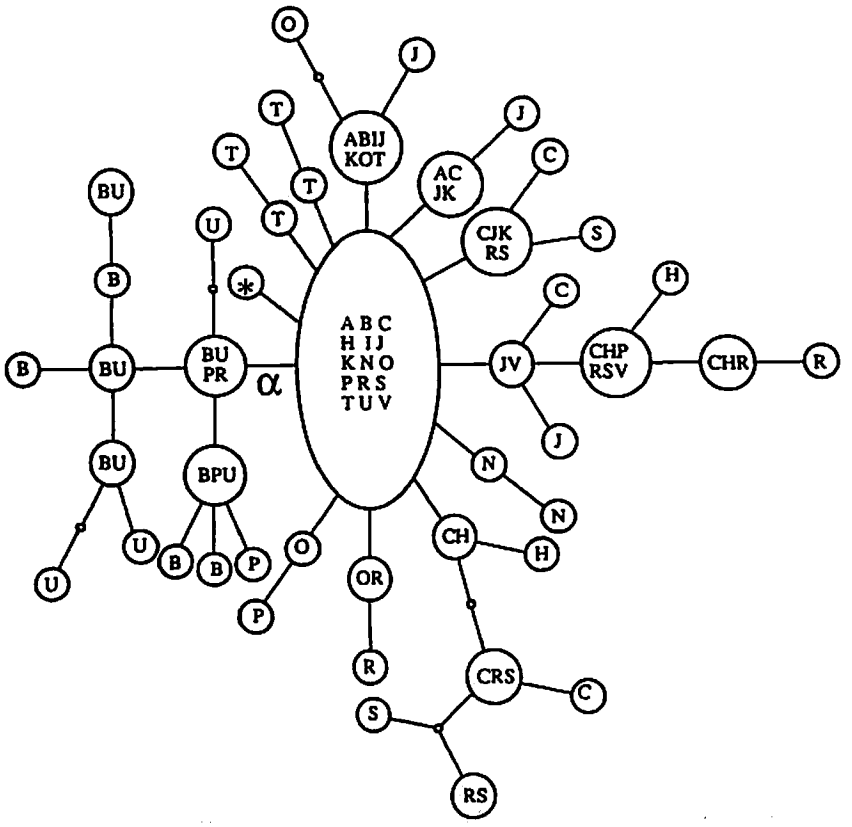

Fig. 18.6. A maximum parsimony tree of 57 human mitochondrial DNA haplotypes found in 15 human populations (modified from Saitou & Harihara, 1996). Letters denote the populations in which each mitochondrial DNA haplotypes were found. Abbreviations of the populations are; A: Ainu, B: Bantu, C: Caucasian, H: Jewish, I: Amerindian, J: Japanese, K: Korean, N: Negritos, O: Oriental, P: Arab, R: Roman, S: Sardinian, T: Tharu, U: Bushman, and V: Vedda. A circle with an asterisk represents 12 different mtDNA types that are one site apart from mtDNA type 1. Small open circles designate intermediate mtDNA types not found but necessary to explain the relationship of known mtDNA types.

estimated to be 0.154/156 = 0.00329. This is equivalent to the 'sequence divergence' of 0.66% (= 0.00329 × 2 × 100); this is not very different from the corresponding value (0.57%) estimated by Cann *et al.* (1987).

The situation in which the root of a gene tree exists in the most common type or allele is not restricted to mtDNA types, but is a general pattern of genealogy for closely related genes. Fig. 18.7(*a*) shows a hypothetical gene genealogy for 14 genes with the common allele (C) and six variant alleles (V1–V6). Because only 7 mutational events were extracted (designated as full circles) and no mutation was observed along the lineages to all the common allele genes, the ancestral gene (the root) is identical with the

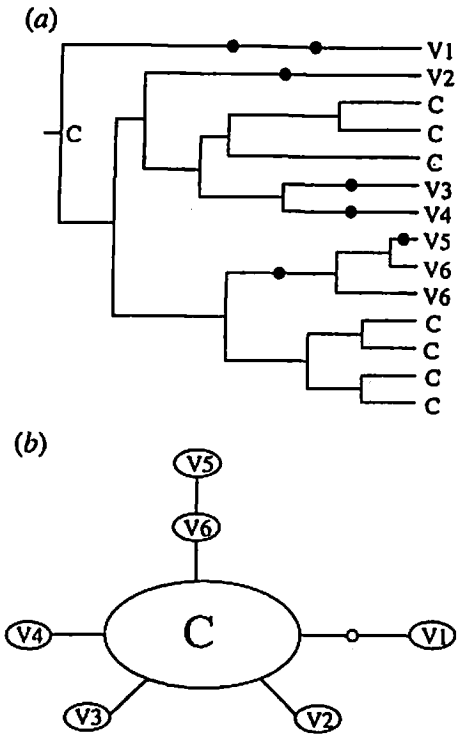

Fig. 18.7. (*a*) A schematic expected tree for 14 genes. Full circles denote mutational events. (*b*) An unrooted gene tree reconstructed by using the observed mutations as shown in the tree above.

common allele. When we reconstruct the gene tree, even the ideal reconstruction (see Fig. 18.7(*b*)) is a gross simplification of the real genealogy. Although the seven C genes are not monophyletic (see Fig. 18.7(*a*)), this cannot be extracted from the reconstructed gene tree of Fig. 18.7(*b*). We should, therefore, be careful in interpreting the branching pattern of a gene genealogy.

We estimated genetic distances among the 15 human populations based on information about the number of nucleotide differences between all the possible pairs of mtDNA types and the frequency of each mtDNA type (Harihara & Saitou, 1989). Fig. 18.8 is an unrooted population tree constructed from that distance matrix data using the neighbour-joining method (Saitou & Nei, 1987). There are some branches with negative lengths in that tree, designated as broken lines. Although this is annoying, the appearance of negative branches is inevitable for non-metric measures such as this genetic distance. This is because the triangle inequality is sometimes violated when there are a number of parallel changes in the allele frequency. For example, let us consider three hypothetical populations (α, β, and γ) and assume that the observed distances were 0.03, 0.05, and 0.01

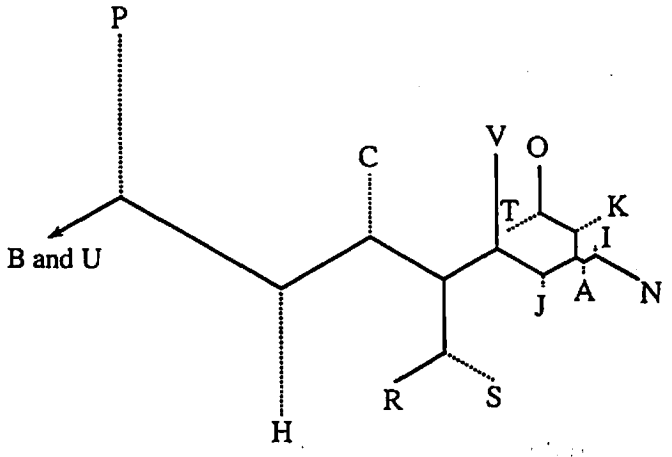

Fig. 18.8. An unrooted neighbour-joining tree for 15 human populations (modified from Saitou & Harihara, 1996). Abbreviations of the populations are the same as those of Fig. 18.6.

for population pairs α–β, β–γ, and γ–α, respectively. In this example, the triangle inequality is violated, for the distance (0.05) between β and γ is larger than the sum (0.04 = 0.03 + 0.01) of distances for β–α and α–γ. The estimates of three branch lengths between the internal node X and three populations then become −0.005, 0.035, and 0.015 for branch αX, βX, and γX, respectively. In fact, branch αX was estimated to be negative.

In any case, let us examine this population tree. Two African populations (Bantu and Bushman) are far apart from the remaining 13 non-African populations. This is clearly because of the distinct clustering of African mtDNA types observed in the gene tree (see Fig. 18.6). Seven *circum*-Pacific populations (Ainu, Amerindian, Japanese, Korean, Negritos, Oriental, and Tharu) are tightly clustered, and the Vedda of Sri Lanka are clustered with these *circum*-Pacific populations. Middle Eastern populations (Arabians and Jews) are located between African populations and the remaining populations. Because the genetic distance data used for producing this population tree were based on the single locus (mtDNA), there are large standard errors in the distance estimates, and the resulting tree may not be completely reliable. We need to study many independent genetic loci in nuclear DNA.

Ancient mitochondrial DNA

Recently so-called ancient DNA studies are have become popular and many ancient human mtDNA sequences have now been published. I will briefly discuss a new aspect of utilizing this ancient DNA data combined

Table 18.1. *Association between burial style and genetic relationship at the Takuta–Nishibun site of Kyushu, Japan*

Burial type	*Number of individuals in mtDNA type*		
	A	non-A	Total
Kamekan (buried in jar-coffin)	6	3	9
Dokoubo (direct burial)	3	14	17
Total	9	17	26

From Oota *et al.*, 1995.

with archaeological data. Oota *et al.* (1995) extracted and amplified a part of the mtDNA D loop region for 26 human bones and teeth found from an archaeological site in the southern part of Japan, dated at *ca.* 2000 BP. Two regions of nucleotide sequences were determined, and phylogenetic trees were constructed (not shown). Nine individuals belonged to the most frequent mtDNA sequence (type A), and the remaining 17 individuals belonged to the other 10 mtDNA sequences, in which 7 of them radiated directly from type A.

There were two types of burial at the site: Kamekan (burial in earthenware jar-coffins) and Dokoubo (direct burial in the earth). To investigate the possibility of a correlation between burial style and genetic relatedness, we computed the probability of obtaining the observed frequency distribution (see Table 18.1). The resultant probability was 0.028 (Fisher's exact test), thus the null hypothesis (of no association) was rejected at the 5% level. This raised two hypotheses about the relationship between burial style and mtDNA type. One assumes that the two burial styles were used at the same period, and the people at the site were buried according to their genetic background (probably kinship). The other hypothesis assumes that these two burial styles were used at different periods, and the genetic constitution of the populations might have been somewhat different between the different periods. This implies an inflow of people with a different genetic background, together with a different culture at least in relation to burial style.

Although ancient DNA data are often used for phylogenetic reconstruction, comparison with archaeological evidence is an important field, as explained above. In this respect, Kurosaki, Matsushita & Ueda (1993) amplified not only mtDNA but also nuclear DNA (short-VNTR loci) from bones of two female individuals (sexing was morphologically determined) buried about 2000 years ago in Japan. These females (mature and juvenile) were buried side by side on the same hill, and both had about 20 cone-shell bracelets on their arms. Because of these characteristics, some archaeol-

ogists considered that they were members of the same family, probably mother and daughter who had ruled over the area as shaman or leader. Kurosaki *et al.* (1993) clearly showed, however, that these females were not genetically related. This finding was contrary to the archaeological conjecture.

Gene trees for nuclear DNA

There are few studies on the reconstruction of gene trees for nuclear DNA, with the exception of the β-globin gene cluster. DNA variation of this region has been extensively studied using restriction enzymes (e.g. Chen *et al.* 1990), and recently Fullerton *et al.* (1994) determined 3-kilobase β-globin sequences for 72 chromosomes. This kind of sequence data will become the standard for future studies of nuclear DNA variation.

Thanks to its extremely high mutation rate and high genetic variation, examination of microsatellite (short-VNTR or STR) loci is becoming popular in human population studies. Bowcock *et al.* (1994) examined 30 microsatellite loci for 148 individuals from 14 human populations, and constructed a colourful tree of 'individuals'. It is not clear from the text whether Bowcock *et al.* considered that tree to be a gene tree. Although a tree of 'individuals' is equivalent to a tree of 'genes' in the case of mtDNA, this does not apply to nuclear DNA data when unlinked loci are used. Since dozens of unlinked loci were used in Bowcock *et al*'s tree of 'individuals', that tree should not be considered as a gene tree in the usual sense. In reality, it presents the relationship of different combinations of unlinked alleles, not the genealogy of individuals. Long branches to extant 'individuals' in the tree of Bowcock *et al.* (1994) do not, therefore, mean long evolutionary times but are merely a reflection of recombination events, and it is erroneous to put a time scale to such a tree.

Population trees for nuclear genes

In contrast to the relatively few studies on nuclear gene trees, those on population trees based on nuclear gene data are abundant. Edwards and Cavalli-Sforza (1964) pioneered the construction of phylogenetic trees of human populations using allele frequency data. Nei and Roychoudhury (1974) estimated the divergence of three major races, and Negroid (African) was estimated to diverge first. This may be the first indication of the 'Out-of-Africa' hypothesis from genetic data.

Saitou, Tokunaga & Omoto (1992) applied the neighbour-joining method for the first time to genetic distance data. Recently, Nei and Roychoudhury (1993) and Bowcock *et al.* (1994) both used the neighbour-

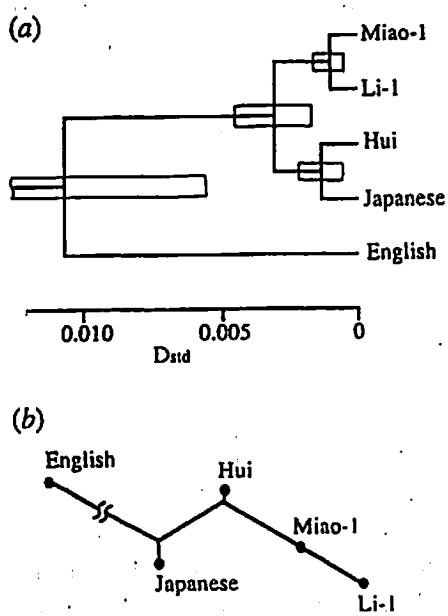

Fig. 18.9. (*a*) A rooted UPGMA tree and (*b*) an unrooted neighbour-joining tree for six human populations (from Saitou *et al.*, 1994).

joining method for reconstructing human population trees. Although they used different datasets (classical markers and microsatellite loci, respectively), similar relationships were obtained.

Theoretically, there is no qualitative difference between a species tree and a population tree. Because a population tree usually means the relationship between populations within a species, however, there is always a chance for intraspecific populations to have high gene flow with each other. Therefore, a rooted tree, in which populations are always assumed to differentiate, may be misleading. In this sense, an unrooted tree representation is more appropriate. Fig. 18.9 shows rooted and unrooted trees for the same genetic distance data of five populations (Saitou *et al.*, 1994). Although Hui, a Muslim population at Hainan Island, is clustered with Japanese in the UPGMA rooted tree (Fig. 18.9(*a*)), it is located between Japanese and Miao-1, another population on Hainan Island. This unrooted population tree suggests that there has been some gene flow between Hui and the surrounding populations of Hainan Island.

Saitou (1995*a*) examined allele frequency data from 12 polymorphic nuclear loci for 30 human populations and constructed an unrooted tree by using the neighbour-joining method (Fig. 18.10). Current human populations are more or less clustered according to their geographical locations; Africa, West and East Eurasia, North and South America, and Sahul land. Sahul land, or the Sahul shelf, existed until about 10 000 years ago, and later

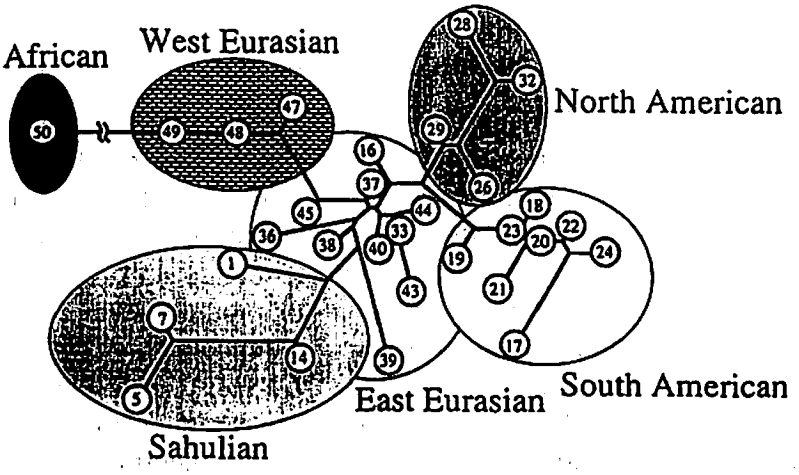

Fig. 18.10. An unrooted neighbour-joining tree for 30 human populations (modified from Saitou, 1995*a*). Population IDs are as follows. Sahulian: 1 = Australian Aborigines (Northern Territory), 5 = Papua New Guinean (North Central Highland), 7 = Papua New Guinean (East Highland), 14 = Micronesian (East Caroline Island). South American: 17 = Yanomama 18 = Makiritare, 19 = Aymara, 20 = Baniwa, 21 = Cayapo, 22 = Macushi, 23 = Wapishana, 24 = Ticuna. North American: 26 = Eskimos (North Alaska), 28 = Athabaskan Indian, 29 = Eskimos (Canada), 32 = Dogrib Indian. East Eurasian: 16 = Polynesian (Samoa Island, now living in New Zealand), 33 = Japanese, 36 = Balinese, 37 = 'Mais, 38 = Filipino, 39 = Negritos, 40 = Han, Northern China, 43 = Ainu, 44 = Korean, 45 = Nepali. West Eurasian: 47 = Indian (South India), 48 = Iranian, 49 = English. African: 50 = Yoruba (Nigeria).

separated into Australia and Papua New Guinea. This continent-wide clustering apparently reflects the history of human population dispersal in the last 100 000 years. Although this pattern is somewhat blurred because of the great human movements particularly within the last 10 000 years, we can still extract the ancient course of human dispersal using genetic data from current populations. Saitou (1995*a*) thus proposed a new classification of human populations based on this genetic affinity tree, as shown in Fig. 18.10. It should be noted that the classification was not meant for the current human populations. It was for those at around the end of Pleistocene, i.e. *ca.* 10 000 BP. The great movement of Polynesian people occurred much later, and the centre of the Pacific was not yet populated at that time. Thus the four clusters surrounding the Pacific (East Eurasian, Sahulian, North American, and South American) can be further grouped to form a '*circum*-Pacific' supercluster. This supercluster corresponds to the 'pan-Mongoloid' cluster of Saitou *et al.* (1992).

Importance of finiteness for evolutionary studies

As the real world is always finite, the course of evolutionary history should also be treated in this finite framework. Random genetic drift caused by the finiteness of the population size is a good example. I would like to emphasize three other aspects of the evolutionary process in which finiteness should be taken into account.

The first is the number of ancestors for one individual. There are 2^n ancestors for a diploid organism such as humans when we go back n generations. This number exceeds the current world population (*ca.* 5×10^9) when $n = 30$ or larger. Of course, the number of individuals at that time (about 6000 years ago if we consider one generation to be 20 years) must have been much smaller than the current level, and inevitably there are many redundancies among the ancestors, i.e. inbreeding. Let us look at this parent–offspring relationship from a different point of view. It is clear that the number of ancestors for a particular mitochondrial DNA is always one, for the circular molecule is inherited without recombination. It immediately follows that the number of ancestral individuals who actually contributed a part of their genetic material is the number of non-recombining units in the genome. Unfortunately, we do not know this number at present. However, the upper limit is the number of nucleotides for a genome, and this is about 3×10^9 for the human genome. Therefore, if the total number of ancestors exceeds this number in a certain generation, there will be ancestral individuals who did not contribute to the genetic composition of a particular present-day descendant. Let us call this ancestor a 'null' ancestor. For example, all male ancestors are null when we consider mtDNA.

The second aspect of finiteness is the number of 'genes' in a genome. Probably the best current estimate of the total number of genes in the human genome is *ca.* 60 000–70 000 (Fields *et al.*, 1994). All the genes from those with housekeeping activities to those involved in complicated ethological characters are in this finite set. If we consider the yet unknown enzymes and proteins expressed in various tissues, the total number of typical genes responsible for biochemical pathways may easily exceed 10 000. Therefore, it is possible that many morphological characters attributed to hereditary factors may be non-hereditary. The same applies to the complicated nature of human brain functions. Unless some unexpected structures that were previously considered to be merely junk are found to be functional, we may be able to map all the functions of genes in the human genome.

The third finiteness is in the number of nucleotides in a non-recombining unit. In theory, the number can be infinite, and it is better to have longer

sequences for obtaining better reconstruction of gene trees from a statistical viewpoint (see e.g. Saitou & Nei, 1986). However, there is always a limit to growth. Horai *et al.* (1995) compared the entire 16.5 kilobase mtDNA genomes of five hominoid species (human, common chimpanzee, pygmy chimpanzee, gorilla, and orang-utan). There will be no need for more study of the mtDNA gene tree of these species, except for intraspecific variation.

Many people are often interested in evolution of a particular gene. In this case, the possible number of nucleotides to be compared is usually much smaller than the entire mtDNA genome. Specific nucleotide changes responsible for the creation or loss of gene function may be delineated, but the estimation of the time frame can be difficult. When the evolutionary time is expected to be quite large, it will be almost impossible to estimate the divergence time of two remotely related genes. In any case, we should be cautious in reconstructing gene trees because of this finiteness.

Acknowledgements

This study was partially supported by a Grant-in-Aid for Scientific Research on Priority Areas (Molecular Evolution) of Ministry of Education, Science and Culture (Japan).

References

Bandelt, H. -J. (1994). Phylogenetic networks. *Verhandlungen des Naturwissenschaftlichen Vereins in Hamburg (NF)*, 34, 51–71.

Bowcock, A. E., Ruiz-Linares, A., Tomfohrde, J., Minch, E., Kidd, J. R. & Cavalli-Sforza, L. L. (1994). High resolution of human evolutionary trees with polymorphic microsatellites. *Nature*, 368, 455–7.

Cann, R. L., Stoneking, M. & Wilson A. C. (1987). Mitochondrial DNA and human evolution. *Nature*, 325, 31–6.

Chen, L. Z., Easteal, S., Board, P. G. & Kirk, R. L. (1990). Evolution of beta-globin haplotypes in human populations. *Molecular Biology and Evolution*, 7, 423–37.

Edwards, A. W. F. & Cavalli-Sforza, L. L. (1964). Reconstruction of evolutionary trees. *Systematics Association Publication*, 6, 67–76.

Farris, S. J. (1972). Estimating phylogenetic trees from distance matrices. *American Naturalist*, 106, 645–68.

Fields, C., Adams, M. D., White, O. & Venter J. C. (1994). How many genes in the human genome? *Nature Genetics*, 7, 345–6.

Fitch, W. M. (1977). On the problem of discovering the most parsimonious tree. *American Naturalist*, 111, 223–57.

Fullerton, S. M., Harding, R. M., Boyce, A. J. & Clegg, J. B. (1994). Molecular and population genetic analysis of allelic sequence diversity at the human β-globin locus. *Proceedings of the National Academy of Sciences, USA*, 91, 1805–9.

Harihara, S. & Saitou, N. (1989). A phylogenetic analysis of human mitochondrial DNA data. *Journal of the Anthropological Society of Nippon*, 97, 483–92.

Hedges, S. B., Kumar, S., Tamura, K. & Stoneking, M. (1991). Human origins and analysis of mitochondrial DNA sequences. *Science*, 255, 737–9.

Horai, S., Hayasaka, K., Kondo, R., Tsugane, K. & Takahata, N. (1995). Recent African origin of modern humans revealed by complete sequences of hominoid mitochondrial DNAs. *Proceedings of National Academy of Sciences, USA*, 92, 532–6.

Kimura, M. (1983). *The Neutral Theory of Molecular Evolution*. Cambridge: Cambridge University Press.

Kurosaki, K., Matsushita, T. & Ueda, S. (1993). Individual DNA identification from ancient human remains. *American Journal of Human Genetics*, 53, 638–43.

Maddison, D. R., Ruvolo, M. & Swofford, D. L. (1992). Geographic origins of human mitochondrial DNA: phylogenetic evidence from control region sequences. *Systematic Biology*, 41, 111–24.

Nei, M. (1987). *Molecular Evolutionary Genetics*. New York: Columbia University Press.

Nei, M. & Roychoudhury, A. K. (1974). Genic variation within and between the three major races of man, Caucasoids, Negroids, and Mongoloids. *American Journal of Human Genetics*, 26, 434–6.

Nei, M. & Roychoudhury, A. K. (1993). Evolutionary relationships of human populations on a global scale. *Molecular Biology and Evolution*, 10, 927–43.

Nerurkar, V. R., Song, K. -J., Saitou, N., Mallan, R. R. & Yanagihara, R. (1993). Interfamilial and intrafamilial genomic diversity of human T lymphotropic virus type I strains from Papua New Guinea and the Solomon Islands. *Virology*, 196, 506–13.

Oota, H., Saitou, N., Matsushita, T. & Ueda, S. (1995). A genetic study of 2,000-year-old human remains of Japan (Yayoi period) using mitochondrial DNA sequences. *American Journal of Physical Anthropology*, 98, 133–45.

Saitou, N. (1995*a*). A genetic affinity analysis of human populations. *Human Evolution*, 10, 17–33.

Saitou, N. (1995*b*). Methods for building phylogenetic trees of genes and species. In *Molecular Biology: Current Innovations and Future Trends*, ed. H. Griffin and A. Griffin., pp. 115–35. Wymondham: Horizon Scientific Press, in press.

Saitou, N. (1996). Reconstruction of gene trees from sequence data. In *Computer Methods for Macromolecular Sequence Analysis*, ed. R. Doolittle, pp. 427–49. Orlando: Academic Press.

Saitou, N. & Harihara, S. (1995). Gene phylogeny and population phylogeny reconstructed from human mitochondrial DNA data. In *Human Evolution in the Pacific Region*, ed. C. K. Ho, G. Kranz and M. Stoneking, Washington: Washington State University Press, in press.

Saitou, N. & Nei, M. (1986). The number of nucleotides required to determine the branching order of three species, with special reference to the human–chimpanzee–gorilla divergence. *Journal of Molecular Evolution*, 24, 189–204.

Saitou, N. & Nei, M. (1987). The neighbour-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4, 406–25.

Saitou, N., Tokunaga, K. & Omoto, K. (1992). Genetic affinities of human populations. In *Isolation and Migration*, ed. D. F. Roberts, N. Fujiki, and K. Torizuka, pp. 118–129. Cambridge: Cambridge University Press.

Saitou, N., Omoto, K., Du, C. & Du, R. (1994). Population genetic study in Hainan

Island, China. II. Genetic affinity analyses. *Anthropological Science*, **102**, 129–47.

Song K. -J., Nerurkar, V. R., Saitou, N., Lazo, A., Blakeslee, J. R., Miyoshi, M. & Yanagihara, R. (1994). Genetic analysis and molecular phylogeny of simian T-cell lymphotropic virus type I: evidence for independent virus evolution in Asia and Africa. *Virology*, **199**, 56–66.

Vigilant, L., Stoneking, M., Harpending, H., Hawkes, L. & Wilson, A. C. (1991). African populations and the evolution of human mitochondrial DNA. *Science*, **253**, 1503–7.

Watterson, G. A. & Guess, H. A. (1977). Is the most frequent allele the oldest? *Theoretical Population Biology*, **11**, 141–160.

Yanagihara, R. & Garruto, R. (1992). Serological and virological evidence for human T-lymphotropic virus type I infection among the isolated Hagahai of Papua New Guinea. In *Isolation and Migration*, ed. D. F. Roberts, N. Fujiki and K. Torizuka, pp. 143–153. Cambridge: Cambridge University Press.

Yanagihara, R., Saitou, N., Nerurkar, V.R., Song, K. J., Bastian, I., Franchini, G. & Gajdusek, D. C. (1995). Molecular phylogeny and dissemination of human T-cell lymphotrophic virus. *Cellular and Molecular Biology*, **41**, S145–61.