ScienceAsia 30 (2004): **239-245**

Selection of Y-Chromosomal Microsatellites for Phylogenetic Study Among Hilltribes in Northern Thailand Using the Decision Tree Induction Algorithm

Daoroong Kangwanpong,[a,*] Jeerayut Chaijaruwanich,[b] Metawee Srikummool[a] and Jatupol Kampuansai[a]

[a] Genetics and Molecular Biology Unit, Department of Biology, Faculty of Science, Chiang Mai University, Chiang Mai 50202, Thailand.
[b] Department of Computer Science, Chiang Mai University, Chiang Mai 50202, Thailand.

* Corresponding author, E-mail: scidkngw@chiangmai.ac.th

Received 9 Jan 2004
Accepted 13 Jul 2004

ABSTRACT: A computational science approach, based on genetic information and methods, was utilized to calculate the minimum number and select the most suitable microsatellite markers for evaluating the diversity and genetic distance among the hilltribes in Northern Thailand. Selecting from previously available and newly published microsatellites, we studied genetic variation and distance - Fst and $(\delta\mu)^2$, of 4 hilltribe populations in Northern Thailand using 15 universal known loci of Y-chromosomal microsatellites. The Decision Tree Induction algorithm in information theory was used to measure impurity of categorizing populations by the number of tandem repeats at each loci and select the minimum number of microsatellites markers with the most discriminating power. Seven selected microsatellites, 8 unselected microsatellites, the original 15 markers, and another four haplotypes used in different global population studies were then employed to construct the UPGMA trees. To validate our results, the Relative Optimality Score (R value) was calculated from the total branch length of the trees, using the tree constructed from genetic distances obtained from 15 Y-chromosomal microsatellites as reference. The graph between R values and different haplotypes was plotted and compared between Fst and $(\delta\mu)^2$ genetic distances. The results show that the haplotype of 7 selected markers is the most reliable compared to all other haplotypes, while the other 8 nonselected loci may be unnecessary for determining the genetic distance in our study. Using the decision tree induction algorithm we were able to select the microsatellites with three levels of discriminating power for differentiating populations, and thereby could reduce the number of Y-chromosomal microsatellite used to half and still achieve the same information.

KEYWORDS: Y-chromosomal microsatellites, haplotype, genetic distance, decision tree induction.

INTRODUCTION

To reconstruct human evolutionary history, be it for the understanding of human origins at a global level or tracing the origin of specific populations at a regional level, many analyses of different types of genetic markers have been performed. Polymorphisms that are assumed to be selectively neutral are preferred, as their evolution in isolated populations is controlled solely by mutation and drift. The results provide a broad picture of human evolution from a genetic perspective, which complements descriptions based on archaeologic and linguistic sources. Among these, microsatellite polymorphisms or short tandem repeats (STRs) are widely used and intensively analyzed in the human population genetic studies[1]. Tetranucleotide markers, 4 base pair motifs, which are widespread in the human genome, present several technical advantages over markers with shorter repeats. They are highly polymorphic and less prone to polymerase slippage, and thus are more often used[2]. The purpose for which microsatellites are useful or informative is for building up measures of genetic distance through either the Fst or $(\delta\mu)^2$ values at a given geographical level, and their consequent usage in understanding the evolution of populations[3].

The major part of the human Y chromosome (i.e. the long arm) consists of polymorphic sequences which are organized into large interspersed tandem repeated arrays. These sequences do not recombine during meiosis, with the result that Y chromosomes are transferred unchanged from generation to generation establishing paternal lineages. It reflects an inherent property of uniparentally inherited DNA sequences (patrilinear inheritance). Populations and subgroups can be dominated by male founder lineages over a long time-span, which, as long as they exist, can only be modified by mutational events[4,5]. Y-chromosomal microsatellites, because of their non-recombining and uniparental inherited natures, have potential for

studying modern human origins and the differentiation of human populations.

The aim of our study of the Y chromosome in human evolution is to demonstrate the true relationship of modern Ys in a tree, which has a root and dated branchpoints. The relatedness of human populations is estimated by the measurement of the frequencies of different Y types in these populations – the more closely related two populations are, the more similar these frequencies are expected to be[6].

With current technology, microsatellites provide the most cost-effective and, at times, most informative genetic marker. However, until recently most of the human population geneticists have used as many single loci of Y chromosomal microsatellites as possible, put together as sets called haplotypes, to evaluate the diversity and genetic distance among populations in different continents of the world[7,8]. We considered the question of whether we could make our study more cost-effective by using less microsatellite loci to differentiate populations, without having to sacrifice any information. To answer this question we used a computational science approach, precisely the decision tree induction algorithm in information theory[9], to measure the impurity of categorizing populations by a number of tandem repeats at each loci and select the minimum number of microsatellite markers with the most discriminating power for our purpose.

Materials and Methods

Studied Populations

The populations under studied were unrelated male volunteers from 4 hill tribe villages, i.e., 19 Karens from Mae Hong Son, 14 Ahkas and 11 Lisus from Chiang Rai and 7 Hmongs from Chiang Mai. Information on linguistic and cultural aspects, village and individual history was obtained by interview.

Blood Sampling and DNA Extraction

Five milliliters of peripheral blood was collected from each individual using a vacutainer coated with EDTA as anticoagulant - . Total genomic DNA was extracted from whole blood samples according to a standard inorganic salting out protocol[10].

Detection of Genetic Variation

Fifteen Y-chromosomal microsatellite loci[11-14] (DYS-19, 388, 389I, 389II, 390, 391, 392, 393, 426, 436, 437, 439 and Y-GATA-A7.1, A7.2, A7.10) were amplified. For each 25 μl PCR volume, 100 ng of total DNA, 800 μM of dNTPs, 3.5 mM $MgCl_2$, 0.1 μM each of the two fluorescent labeled oligonucleotide primers, 2.5 μl of 10X PCR buffer and 0.35 units of AmpliTag Gold™ Polymerase (Perkin-Elmer) were thermal cycled. PCR cycling conditions for these loci were: an initial denaturation step of 95°C for 10 minutes, followed by 30 cycles of 94°C for 1 minute, 54°C for 1 minute, and 72°C for 1 minute, and a final extension step at 72°C for 7 minutes. Amplified DNA fragments were automatically detected with an ABI 377 automated sequencer using GENESCAN™ and GENOTYPER™ software (Applied Biosystems).

Classification by Decision Tree Induction

The basic algorithm for decision tree induction (DTI) is a greedy algorithm that constructs a decision tree in a top-down recursive divide-and-conquer manner[6]. The algorithm is summarized as follows.

Algorithm Generate_Decision_Tree (S, M)

(Generate a decision tree from the given subjects S and microsatellites M)

Input: S: set of subjects, each subject is represented by set of alleles

M: set of candidate microsatellites (Fig 1)

Output: the decision tree

1) create a node N
2) if alleles are all of the same tribe, C then
3) return N as a leaf node labeled with the population C
4) if set of candidate microsatellites is empty then
5) return N as leaf node labeled with the most common population in alleles // majority voting
6) select the most discriminating microsatellite, the microsatellite among the set of candidate microsatellites with the highest information gain
7) label node N with the most discriminating microsatellite
8) for each known repeat value a_i of the most discriminating microsatellite // partition alleles
9) grow a branch from node N for the condition the most discriminating microsatellite = a_i
10) let s_i be the set of alleles for which the most discriminating microsatellite = a_i // a partition
11) if s_i is not empty then
12) attach the node returned by Generate_Decision_Tree (s_i, M/{most discriminating microsatellite})

The basic idea of the algorithm is as follows. The tree starts as a single node representing the given alleles (step 1). If the alleles are all of the same population, then the node becomes a leaf and is labeled with that population (steps 2-3). If there are no remaining microsatellites, majority population voting is applied (steps 4-5). Otherwise, the algorithm uses an entropy-based measure known as information gain as a heuristic for selecting the microsatellite that will best separate the alleles into individual populations (step 6). This microsatellite becomes the 'test' or 'decision' microsatellite at the node (step 7). A branch is created

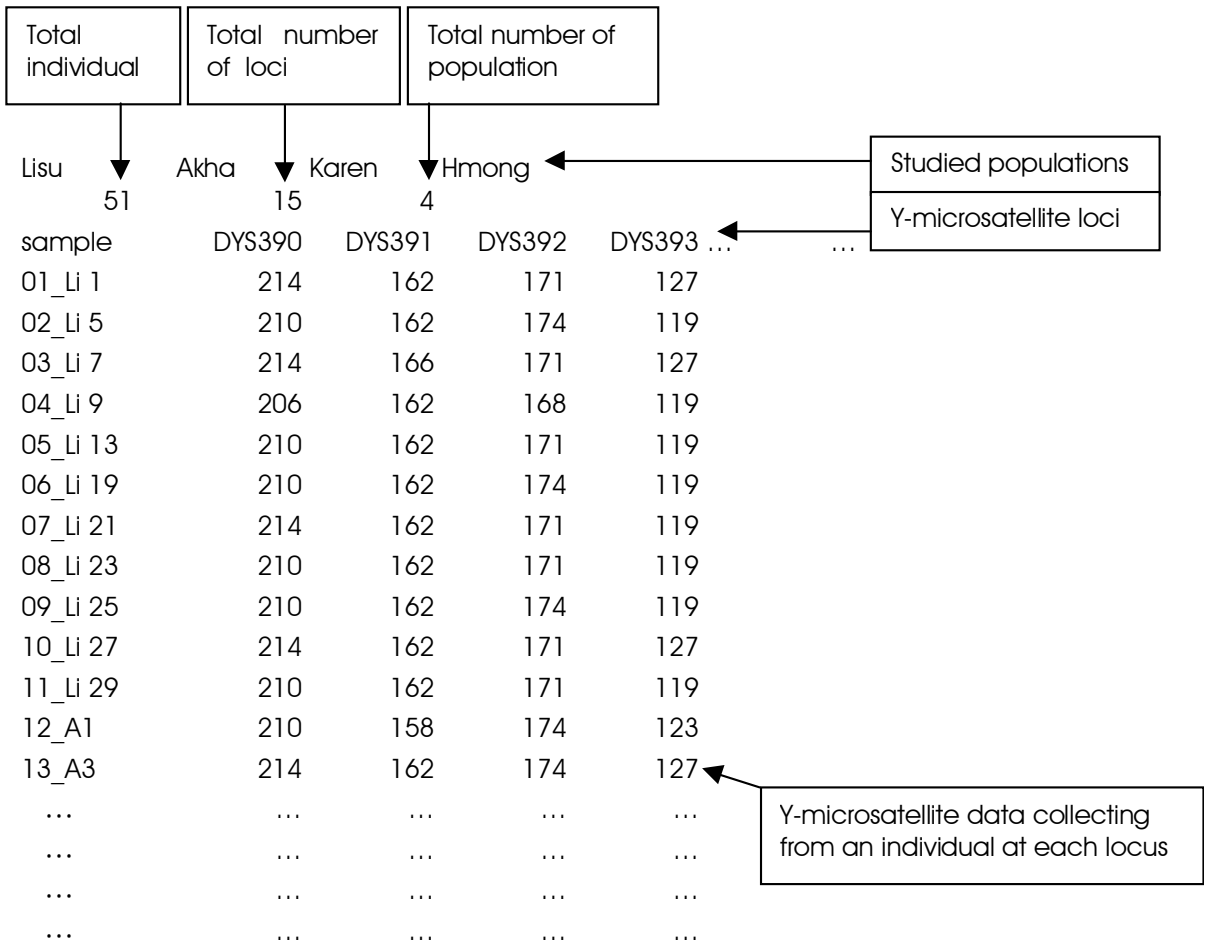

Fig 1. Input data format for Decision Tree Induction.

for each known repeat value of the test microsatellite, and the alleles are partitioned accordingly (steps 8-10). The algorithm uses the same process recursively to form a decision tree for the alleles at each partition (steps 11-12).

Microsatellite Selection Measure

The information gain measure is used to select the most discriminating microsatellite at each node in the tree. Such a measure is referred to as a microsatellite selection measure or a measure of the goodness of split. The microsatellite with the highest information gain (or greatest entropy reduction) is chosen as the test microsatellite for the current node. This microsatellite minimizes the information needed to classify the alleles in the resulting partitions and reflects the least randomness or 'impurity' in these partitions. Such an information-theoretic approach minimizes the expected number of tests needed to classify an object and guarantees that a simple (but not necessarily the simplest) tree is found.

Let S be a set consisting of s alleles. Suppose alleles are collected from m distinct populations, C_i (for i = 1, …, m). Let s_i be the number of alleles of S in population C_i. The expected information needed to classify a given allele is given by

$$I(s_1, s_2, ..., s_m) = -\sum_{i=1}^{m} p_i \log_2 p_i$$

where p_i is the probability that an arbitrary allele belongs to population C_i and is estimated by s_i/s. Note that a log function to the base 2 is used since the information is encoded in bits.

Let microsatellite A have v distinct repeat values, $\{a_1, a_2, \ldots, a_v\}$. Microsatellite A can be used to partition S into v subsets, $\{S_1, S_2, \ldots, S_v\}$, where S_j contains those alleles in S that have value a_j of A. If A was selected as the most discriminating microsatellite (i.e. the best microsatellite for splitting), then these subsets would correspond to the branches grown from the node containing the set S. Let s_{ij} be the number of alleles of population C_i in a subset S_j. The entropy, or expected information based on the partitioning into subsets by A, is given by

$$E(A) = \sum_{j=1}^{v} \frac{s_{1j} + \ldots + s_{mj}}{s} I(s_{1j}, \ldots, s_{mj})$$

The term $\frac{s_{1j} + \ldots + s_{mj}}{s}$ acts as the weight of the j^{th} subset and is the number of alleles in the subset (i.e., having value a_j of A) divided by the total number of alleles in S. The smaller the entropy value, the greater the purity of the subset partitions. Note that for a given subset S_j,

$$I(s_{1j}, s_{2j}, \ldots, s_{mj}) = -\sum_{i=1}^{m} p_{ij} \log_2 p_{ij}$$

where $p_{ij} = s_{ij} / |S_j|$ and is the probability that an allele in S_j belongs to tribe C_i.

The encoding information that would be gained by branching on A is

$$Gain(A) = I(s_1, s_2, \ldots, s_m) - E(A)$$

In other words, Gain (A) is the expected reduction in entropy caused by knowing the repeat value of microsatellite A.

The algorithm computes the information gain of each microsatellite. The microsatellite with the highest information gain is chosen as the most discriminating microsatellite, branches are created for each repeat value of the microsatellite, and the alleles are partitioned accordingly.

Genetic Analyses

1. Pairwise Fst and $(\delta\mu)^2$ genetic distances between populations were calculated using Microsat (available at http://lotka.stanford.edu/microsat.html) from different haplotypes (Table 1). UPGMA trees were constructed from both distances using the MEGA 2 program (available at http://www.megasoftware.net)

Data used for genetic distance calculation and UPGMA tree construction were allelic frequencies investigated at different combined loci (haplotypes) listed in Table 1.

Sets 1, 2 and 3 were from the present study. Set 1 was the haplotype of all 15 loci used (reference set) while set 2 and 3 were haplotypes of DTI selected and unselected loci, respectively. The other four sets were frequently used haplotypes from different groups of population geneticists.

2. For accuracy and statistical tests of UPGMA trees, we used the relative optimality score (R) defined below[6] to compare total lengths between UPGMA trees constructed from different Fst and $(\delta\mu)^2$ genetic distances.

$$R = (TL - TL_C) / TL_C$$

when TL is the total length (sum of branch lengths) of UPGMA trees constructed from data of set 2 to 7 haplotypes and TL_C is the sum of branch lengths of UPGMA trees constructed from data of set 1 haplotype - our ideal tree.

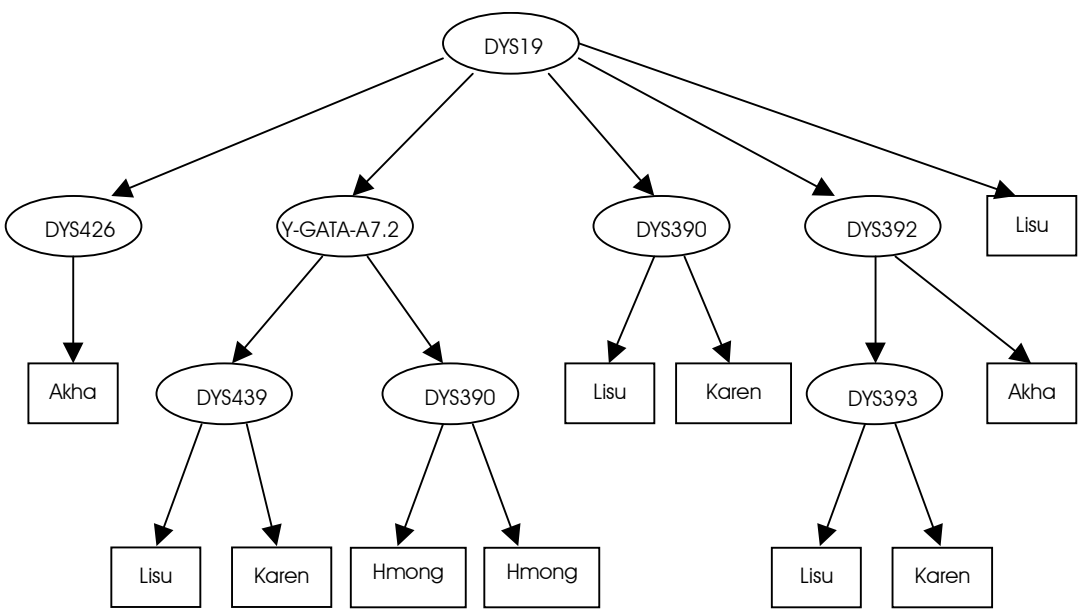

Fig 2. Decision tree of Y-chromosomal microsatellites.

Table 1. Different haplotypes used for validation of our results.

Set	Haplotype (loci combination)	Population	Reference
1	DYS-19, 388, 389I, 389II, 390, 391, 392, 393, 426, 436, 437, 439 and Y-GATA-A7.1, A7.2, A7.10	present study	in materials and methods
2	DIC selected haplotype	present study	from results
3	Non-selected haplotype	present study	from results
4	DYS19-388-390-391-392-393	Southwest Asia	15
		British Isles	16
5	DYS19-389I-389II-390-391-392-393	Basques, Iberia	17
6	DYS19-388-389I-390-391-392-393	Columbia	18
		Finland	19
7	DYS19-388-389I-389II-390-391-392-393	Central Asia	20

Table 2. TL and R values comparing between each haplotype set and the reference set 1.

Haplotype set	reference 1		selected 2		non-selected 3		different studies 4		5		6		7	
Distance	Fst	$(\delta\mu)^2$	Fst	$(\delta\mu)^2$	Fst	$(\delta\mu)^2$	Fst	$(\delta\mu)^2$	Fst	$(\delta\mu)^2$	Fst	$(\delta\mu)^2$	Fst	$(\delta\mu)^2$
TL	0.986	0.859	1.119	0.963	0.878	0.750	1.305	0.914	1.317	0.899	1.217	0.888	1.226	0.895
R	0.000	0.000	0.135	0.121	**-0.11**	**-0.13**	0.324	0.063	0.336	0.046	0.235	0.034	0.244	0.041

Results

The decision tree induction algorithm described previously provides the decision tree shown by Fig 2. It identifies a haplotype of 7 loci namely, DYS19, DYS390, DYS392, DYS393, DYS426, DYS439 and Y-GATA-A7.2 (set 2 in Table 1). DYS19 was recognized to be the most discriminating locus. The discriminating order of the other loci diminish hierarchically into 2 ordered layers, the first with DYS426, Y-GATA-A7.2, DYS390, and DYS392 and the second with DYS439, DYS390, and DYS393. The other 8 loci—DYS388, DYS389I, DYS389II, DYS391, DYS436, DYS437, Y-GATA-A7.2 and A7.10 (set 3 in Table 1) were rejected.

The UPGMA trees were constructed from the data of seven haplotype sets and the relative optimality score (R) was used to prove the validity of the trees. The data from 15 microsatellite loci (set 1 haplotype) were chosen as a reference for constructing our ideal tree, thus the relative optimal score (R) of set 1 is zero. R values of other sets are positive except the non-selected haplotype (set 3). Moreover, we found that R values of

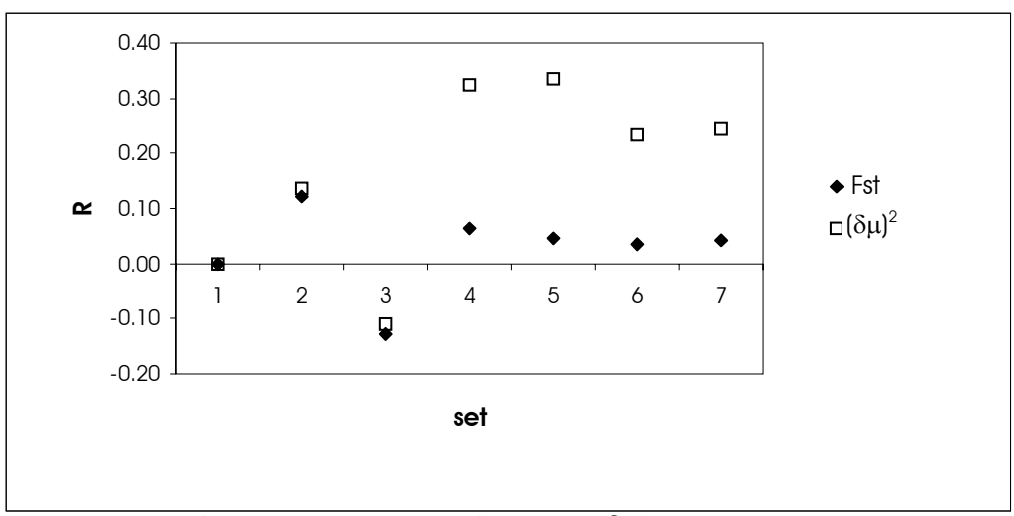

Fig 3. Comparing R values of each haplotype set deriving from Fst and $(\delta\mu)^2$ genetic distances.

sets 2 and 3 were not different when their UPGMA trees were constructed from Fst versus $(\delta\mu)^2$ genetic distances, while those of sets 4 to 7 were different, as shown in Table 2 and Fig 3.

Discussion

Some of Y-microsatellites were proved to be suitable loci in human population studies, such as DYS19 and DYS393, and these loci were commonly used by many researchers. A combination of those proven loci was used to investigate population affinities. The number of loci used in such studies depends on many factors, but about 6 to 8 loci has always been an appropriate number and has been widely used[15-20].

In this study, the commonly used loci, DYS19, 388, 389I, 389II, 390, 391, 392 and 393, were combined with the additional loci DYS426, 436, 437, and 439 and Y-GATA-A7.1, A7.2, and A7.10 to make the reference set of haplotypes for an evaluation. Allele frequency and distribution of every locus from each individual, as shown in Fig 1, was utilized for DTI. Genetic diversity indices among hilltribe populations analyzed by DYS19, 390 and 393 were published elsewhere[21]. The result from DTI showed that DYS19, one of the most extensively analyzed microsatellites in human population genetics[22], was recognized as the most appropriate locus for studying the 4 hill-tribe populations of Thailand. Together with the other two selected orders of old and novel microsatellite loci, DYS 390, 392, 426, and Y-GATA-A7.2 and DYS390, 393, and 439, it made the selected 7 loci set. With different sets of haplotypes we calculated the genetic distances and constructed the UPGMA trees for comparison.

The same general principle with the concept of distance in a geographic sense implies for the genetic distance - the smaller the genetic distance between two populations, the closer they are genetically. A genetic distance is, therefore, simply a measure of how different two populations are genetically. In particular, it can tell which two populations are the most similar in their allele frequency and distribution. For this, we would not want to base any inferences on only one allele, but we would compute the genetic distances for many dozens of alleles and then average them. This requirement leads to high cost and time consumption. In this study, using the decision tree induction algorithm we are able to select the microsatellites with three levels of discriminating power for differentiating populations by which we could reduce the number of Y-chromosomal microsatellite used to half (Fig 2) and still achieve the same information.

Theoretically, for phylogenetic inference it is often assumed that the sample size (n) is large. Practically, however, the number of samples investigated is mostly small, which is subject to stochastic errors. Moreover, to know the reliability of the newly constructed phylogenetic tree, for which the topological errors is the primary concern, is also very important. Therefore, the performance of optimization criteria is recommended. When the true topology of the phylogenetic tree is not known and n is small, the relative optimality score (R), which compares the sum of branch lengths between different trees, is usually used to prove the validity of the tree. R is positive when $TL > TL_C$, zero when $TL = TL_C$ and negative when $TL < TL_C$. The tree will lack of reliability when TL is smaller than TL_C or R is negative. On the other hand, the tree will be valid when R is positive and more reliable when R is close to zero[3].

In our study, we assumed that our ideal tree, which was constructed from the genetic distances calculated from set 1 haplotype data, had a true topology. The evaluation of the branch length errors was then performed using branch lengths of the ideal tree as reference. The validation of the test could be seen by the relative optimality score. The R values of selected haplotype (set 2) were the only positive and concordant values calculating from TC of either the Fst or $(\delta\mu)^2$ genetic distance. The non-selected haplotype (set 3) showed concordant values between Fst and $(\delta\mu)^2$ genetic distances but with negative scores, thus was invalid. The reason of the validation is due to the topology deviation of the UPGMA tree. In practice, the true topology is almost never known, and therefore, the reliability of the topology obtained is usually tested by examining the statistical confidence of various parts of the topology, in our case the sum of the branch lengths.

Acknowledgements

The authors wish to thank all of the DNA donors for their samples and cooperation and to the staff of Tribal Research Institute, Chiang Mai, Thailand, for the field research organization. Laboratory and field expenses were supported with funds from the Department of Biology, Faculty of Science, Chiang Mai University and the Program for Population Genetics, Harvard School of Public Health.

References

1. Ruiz-Linares A (1999) Microsatellites and the reconstruction of the history of human populations. In: *Microsatellites: Evolution and Applications* (Edited by Goldstein DB and Schlötterer C), pp 183-97. Oxford University Press, New York.
2. Eisen JA (1999) Mechanistic basis for microsatellite instability. In: *Microsatellites: Evolution and Applications* (Edited by Goldstein DB and Schlötterer C), pp 34-48. Oxford University Press, New York.

3. Nei M and Kumar S (2000) *Molecular Evolution and Phylogenetics*. pp 231-64. Oxford University Press, New York.
4. Nei M and Roychoudhury AK (1993) Evolutionary and relationships of human populations on a global scale. *Mol Biol Evol* **10**, 927-43.
5. Cavalli-Sforza LL (1997) Genes, peoples and languages. *Proc. Natl Acad Sci (USA)* **94**, 7719-24.
6. Nei M and Kumar S (2000) *Molecular Evolution and Phylogenetics*. pp 165-86. Oxford University Press, New York.
7. Roewer L *et al.* (1992) Simple repeat sequences on the human Y chromosome are equally polymorphic as their autosomal counterparts. *Hum Genet* **89**, 389-94.
8. Mathias N *et al*. (1994) Highly informative compound haplotype for the human Y chromosome. *Hum Mol Genet* **3**, 115-23.
9. Quinlan JR (1993) *C4.5:Programs for Machine Learning*. Morgan Kaufmann Publishers, San Mateo, CA.
10. Seielstad M Bekele E Ibrahim M Toure A and Traore M (1999) A view of modern human origin from Y chromosome microsatellite variation. *Genome Res* **9**, 558-67.
11. Kayser M *et al.* (1997) Evolution of Y-chromosomal STRs: A multicenter study. *Int J Legal Med* **110**, 125-33.
12. Thomas MG Bradman N and Flinn HM (1999) High throughput analysis of 10 microsatellite and 11 diallelic polymorphisms on the human Y-chromosome. *Hum Genet* **105**, 577-81.
13. White PS Tatum OL Deaven LL and Longmire JL (1999) New, male-specific microsatellite markers from the human Y chromosome. *Genomics* **57**, 433-7.
14. Ayup Q Mohyuddin A Qamar R Mazhar K Zerjal T Qasim Mehdi S and Tyler-Smith C (2000) Identification and characterization of novel human Y-chromosomal microsatellites from sequence database information. *Nucleic Acid Research* **28(2)**, e8, i-v.
15. Bandelt HJ Forster P and Rohl A (1999) Median-joining networks for inferring intraspecific phylogenies. *Mol Biol Evol* **16**, 37-48.
16. Wilson JF Weiss DA Richards M Thomas MG Bradman N and Goldstein DB. (2001) Genetic evidence for different male and female roles during cultural transition in the British Isles. *Proc. Natl Acad Sci (USA)* **98**, 5078-83.
17. Gonzalez-Neira A Gusmão L Brión M Lareu MV Amorim A and Carracedo A. (2000) Distribution of Y-chromosome STR defined haplotypes in Iberia. *Forensic Science International* **110(2)**, 117-26.
18. Ruiz-Linares A *et al* (1999) Microsatellites provide evidence for Y chromosome diversity among the founders of the new world. *Proc. Natl Acad Sci (USA)* **96**, 6312-7.
19. Kittles RA *et al*. (1998) Dual origins of Finns revealed by Y chromosome haplotype variation. *Am J Hum Genet* **62**, 1171-9.
20. Perez-Lezaun A *et al*. (1999) Sex-specific migration patterns in Asian populations, revealed by analysis of Y-chromosome short tandem repeats and mtDNA. *Am J Hum Genet* **65**, 208-19.
21. Srikummool M Kangwanpong D Singh N and Seielstad M (2000) Y-chromosomal variation in uxorilocal and patrilocal populations in Thailand. In: *Genetic, Linguistic and Archaeological Perspectives on Human Diversity in Southeast Asia* (Edited by Jin L Seielstad M and Xiao C), pp 69-82. World Scientific Publishing Co Ltd, Singapore.
22. Quintana-Murci L *et al*. (1999) Y-chromosome specific YCAII, DYS19 and YAP polymorphisms in human populations: a comparative study. *Ann Hum Genet* **63**, 153-66.