JOURNAL OF FORENSIC SCIENCES

J Forensic Sci, January 2011, Vol. 56, No. S1
doi: 10.1111/j.1556-4029.2010.01614.x
Available online at: onlinelibrary.wiley.com

PAPER

CRIMINALISTICS

Aimé Ntwari,[1,†] M.Sc.; Abdellali Kelil,[2,†] Ph.D.; Régen Drouin,[1] Ph.D., M.D.; Ernest Monga,[3] Ph.D.; Shengrui Wang,[4] Ph.D.; Ryszard Brzezinski,[5] Ph.D.; Marc Bronsard,[1] M.Sc.; and Ju Yan,[1] Ph.D.

DNAc: A Clustering Method for Identifying Kinship Relations Between DNA Profiles Using a Novel Similarity Measure*

ABSTRACT: After decades of refinement, DNA testing methods have become essential tools in forensic sciences. They are essentially based on likelihood ratio test principle, which is utilized specifically, by using as prior knowledge the allele frequencies in the population, to confirm or refute a given kinship hypothesis made on two genotypes. This makes these methods ill suited when allele frequencies or kinship hypotheses are unavailable. In this paper, we introduce DNAc, a new clustering methodology for DNA testing based on a new similarity measure that allows an accurate retrieval of the degree of relatedness among two or more genotypes, without relying on kinship hypotheses or allele frequencies in the population. We used DNAc in analyzing microsatellite DNA sequences distributed among 12 genotypes from normal individuals from two distinct families. The results show that DNAc accurately determines kinship among genotypes and further gathers them in the appropriate kinship groups.

KEYWORDS: forensic science, DNA, microsatellite, kinship, clustering, similarity

Although each individual's DNA profile is unique, the DNA profiles of biologically related people show specific kinds of similarity to each other. With DNA testing, these similarities can be detected and used to link related individuals. Existing DNA testing approaches for determining biological relationship target noncoding DNA sequences in the human genome. The polymorphisms, or variant versions of sequences, found in the genome, including restriction fragment length polymorphisms, variable-number tandem repeats, short tandem repeats, single-nucleotide polymorphisms, and microsatellites or simple sequence repeats (1–3), occur in informative patterns and generational shifts in a given family and can thus serve as genetic loci for DNA testing.

[1]Division of Genetics, Department of Paediatrics, Faculty of Medicine and Health Sciences, University of Sherbrooke, 3001 12th Avenue North, Sherbrooke, QC J1H 5N4, Canada.
[2]Michnick Laboratory, Department of Biochemistry, Faculty of Medicine, University of Montreal, 2900, boulevard Édouard-Montpetit, Montreal, QC H3T 1J4, Canada.
[3]Department of Mathematic, Faculty of Science, University of Sherbrooke, 2500, boul. de l'Université, Sherbrooke, QC J1K 2R1, Canada.
[4]ProspectUS Laboratory, Department of Computer Science, Faculty of Science, University of Sherbrooke, 2500, boul. de l'Université Sherbrooke, QC J1K 2R1, Canada.
[5]Molecular Biotechnology Laboratory, Department of Biology, Faculty of Science, University of Sherbrooke, 2500, boul. de l'Université Sherbrooke, QC J1K 2R1, Canada.

*This study was supported by grants from Fonds de la Recherche en Santé du Québec (FRSQ), Fondation des étoiles, Faculté de médecine et des sciences de la santé of Université de Sherbrooke, and Centre de Recherche Clinique Étienne-Le Bel to J. Yan.

†These authors contributed equally.

Received 28 April 2009; and in revised form 28 Nov. 2009; accepted 5 Dec. 2009.

The use of microsatellites as genetic loci for the purpose of DNA testing is one of the most frequently employed methods for establishing biological relationships. This is attributed to their high heterozygosity and polymorphism. To take advantage of their information content, we have chosen to use microsatellites as genetic loci in all of our experiments. However, DNAc can be used with any kind of genetic loci that have high heterozygosity and polymorphism.

Existing approaches for DNA testing are based on the likelihood ratio principle (4,5). They are devised specifically to confirm or refute a given kinship hypothesis between two or more genotypes (5,6), but not to identify unknown kinship relations. In addition, they rely on knowledge of allele frequencies in the population. However, because of the absence of sufficient studies about the distribution of allele frequencies in various populations, this evaluation remains a major challenge.

In this paper, we present DNAc, a new and original methodology for DNA testing. Unlike the likelihood ratio method that is used merely to confirm or refute a given kinship hypothesis, DNAc is a methodology that outputs the kinship degrees of the input genotypes. DNAc comprises a new approach for encoding alleles, a new similarity measure, and a new profile-clustering algorithm. DNAc makes it possible to accurately pinpoint the degree of kinship between two or more individuals without relying on population allele frequencies. DNAc has the further advantage of being effective even with small data sets.

To show the effectiveness of DNAc, we used a test based on polymerase chain reaction (PCR) to investigate different patterns of microsatellite DNA sequences distributed among normal family members. We chose microsatellites with at least 80% heterozygosity (see supplementary data in S1 and S2). The data were then subjected to computer-assisted analysis. Our results clearly

demonstrate the usefulness of DNAc for determining biological relationships among DNA profiles and thus for gathering them into appropriate parental groups.

Materials and Methods

Samples and Loci

Two families were analyzed using 112 and 32 microsatellites for independent loci, respectively. In this work, we used microsatellites as loci, but any kind of loci with high heterozygosity and high polymorphism could be used with DNAc. The first family was comprised of a brother and a sister, a half-brother to both, and an unrelated individual as control. The second family was comprised of a mother, a father, their daughter, and a half-sister of the daughter. The description of all loci is available as supplementary data (Data S1 and S2) with this paper. The loci were chosen for their high percentage of heterozygosity.

PCR Tests

The DNA was extracted from 200 μL of peripheral blood with the QIAamp DNA mini kit (QIAGEN, Mississauga, Ontario). Then the microsatellites were amplified by the use of PCR, using specific sequences flanking the microsatellites as primers (Integrated DNA Technology, Coralville, IA). The procedure consisted of the manufacturer's recommended protocol with a slight modification. Briefly, PCR amplification was carried out in 10 μL of reaction solution containing 10–20 ng DNA, 200 μM of each dNTP, 2.5 mM $MgCL_2$, 0.1 μM of each primer including the labeled m13 queue, and 0.3 units of HotStar Taq (enzyme) (QIAGEN) in 1× PCR buffer and overlaid with oil. Denaturation was performed at 95°C for an interval of 1–3 min, followed by annealing at 55°C for 30 sec, and then chain extension at 72°C for 40 sec, for a total of 35 cycles. To view the results, we loaded the labeled DNA on a 6% polyacrylamide gel in an automatic sequencing system (DNA 4300 sequencer; Li-Cor Company, Lincoln, NE). When the gel was laser scanned for bands, the results often resembled supermarket bar codes (6). We then evaluated microsatellite bands to differentiate parental origin. An example of the visual result is shown in Fig. 1.

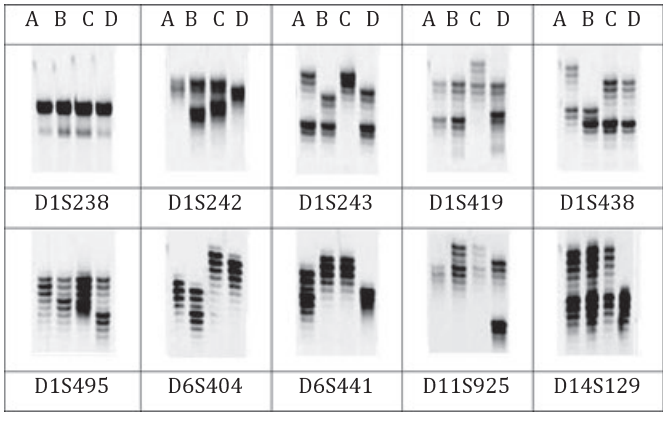

FIG. 1—*Representative results of the laser scans on the polyacrylamide gel obtained from PCR with 10 microsatellite markers. (For each panel, the first three lanes A, B, and C are members of the same family; the last lane D is a control sample from an unrelated individual.) The name of each marker is provided below the panel (i.e., the Ladder marker is not shown on different figures).*

Encoding

The aim of the encoding technique presented in this section is to transform the sizes of the alleles in the genotypes into integers, such that alleles with the same size from the same locus in all the genotypes yield the same integer. The encoding technique is described later. Now, let M be a set of DNA profiles to be studied, including S different alleles, and let E be the set of all possible positions in the electrophoresis gel that each of these S alleles can take. $P_{i,j}$ is the position of the jth allele of the ith DNA profile. We annotate each allele using the algorithm described later.

As an example, the result of the encoding of locus D6S1671 is shown in Fig. 2. This example shows that the way Algorithm 1 assigns a code to each allele. The identification of codes in this example is arranged left to right and top to bottom. At each iteration of Algorithm 1, the position of the current allele is compared to the positions of all previously encoded alleles. If at least one previously encoded allele has the same position as the current allele, then the latter takes the same code; otherwise, the current allele takes the highest code incremented by one.

Encoding Algorithm

Input: DNA profiles
Initialization: E = ø, S = 1
  For i = 1 to M do
  For j = 1 to 2 do
  If ($P_{i,j} \in$ E) then
    $P_{i,j}$ = s
    $S = S + 1$
    $E = E + \{P_{i,j}\}$
  End If
Output: Positions of alleles

The New Similarity Measure

Almost all existing similarity measures used in the DNA field are based on sequences. To the best of our knowledge, we are the

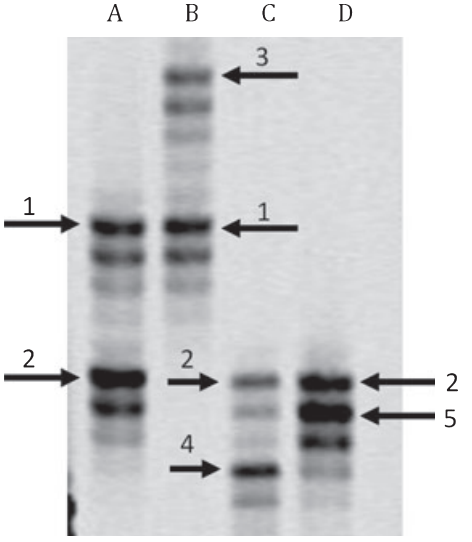

FIG. 2—*The encoding results of Algorithm 1 for one microsatellite marker of four individuals. A is the half-brother of B and C; B is the sister of C; D is an unrelated individual. The numbers 1–5 are the resulting encoding identifications of different alleles.*

first to present a similarity measure that directly handles the number of repetitions expressed by the size of each allele in each locus, obtained by PCR amplification. The kinship between two DNA profiles can be deduced from the proportion of shared genetic material. The proportion can be expressed in terms of similarity. Now, let M be the set of DNA profiles obtained using the DNA extraction and PCR protocol presented earlier. First, we use Algorithm 1 to encode the set of all alleles. Second, we compute the similarity measure between DNA profiles X and Y, using the formula below:

$$S_{X,Y} = \frac{1}{N}\sum_{i=1}^{N} S^i_{X,Y} \quad \text{such that} \quad S^i_{X,Y} = \begin{cases} \frac{m^i_c}{m^i_u - m^i_c} & \text{if} \quad m^i_u - m^i_c \neq 0 \\ 1 & \text{if} \quad m^i_u - m^i_c = 0 \end{cases}$$

where, N is the number of loci, $S^i_{X,Y}$ is the similarity, m^i_c is the number of common alleles, and m^i_u is the total number of different alleles for a given locus ith of X and Y together. In real life, similarities between different genotypes can arise by chance. To reduce the effect of similarities of this kind on the DNAc accuracy, we compute the similarity measure between genotypes after discarding the noninformative loci (i.e., those present in all the DNA profiles used in the same test). Table 1 gives an example of how similarity is computed for the profiles shown in Fig. 2.

Relationship Testing

Paternity Testing—The DNA profile of a child is inherited from both parents, with one allele for each locus derived from the father's DNA profile and the other from the mother's profile. Thus, the exclusion of paternity will be ascertained if at least one locus from the child does not share any allele with the same locus from the potential father. A sufficient condition for excluding paternity with a probability equal to 1 is:

$$P_{X,Y} = \prod_{i=1}^{N} S^i_{X,Y} = 0$$

For the case where the product aforementioned is nonzero, we define the events A, B, and A' (in a probabilistic sense) such that:

- Event A: X is the father of Y
- Event B: Y has the same genetic locus as X
- Event A': X is not the father of Y

We have the following probabilities:

$P(A|B)$ = The probability that X is the father of Y given the observed genetic profiles.

$P(B|A)$ = The probability of the observed genetic profiles given that X is the father of Y.

$P(B|A')$ = The probability of the observed genetic profiles given that X is not the father of Y (equal to the square of the frequency of the genetic locus in the population).

$P(A)$ = The assumed probability before testing that X is the father of Y.

$P(A')$ = The assumed probability before testing that X is not the father of Y.

$P(A|B)$ is obtained by Bayes' formula, as follows:

$$P\left(\frac{A}{B}\right) = \frac{P\left(\frac{B}{A}\right)P(A)}{P\left(\frac{B}{A}\right)P(A) + P\left(\frac{B}{A'}\right)P(A')} \quad \text{where} \quad P\left(\frac{B}{A}\right) = \prod_{i=1}^{N} S^i_{X,Y}$$

Because there is a half probability that a child will inherit a specific genetic allele locus from a man if he is the child's biological father, here, only $P(A)$, and $P(A')$, and $P(B|A')$ need to be identified before plugging the values into Bayes' formula as $P(A') = 1.0 - P(A)$. To evaluate $P\left(\frac{B}{A'}\right)$, we need the frequency of each allele in the whole population from which the profiles are derived. This assumes a survey of the allele's frequency in the population has already been conducted, which is not always the case. All that remains is to identify $P(A)$. Remember that $P(A)$ is the probability, assumed prior to testing, that the man is the biological father of the child. In paternity testing, this probability is often assumed to be 50%. For 10 loci of a given frequency f_i, we can see that the probability of paternity is a function that decreases with f_i (see Table 2 for details). An informative locus would be one occurring with low frequency; otherwise, we would need to use a high number of loci.

Other Relationship Testing—The closeness of any relationship by biological inheritance is analyzable by this technique. To measure the degree of kinship, we have to compute the percentage of common genetic inheritance. For a kinship of degree d, it is well known that the proportion of genetic inheritance is given by 0.5^d. Thus, for a given similarity measure $S_{X,Y}$ between two DNA profiles X and Y, the degree d of kinship is computed simply as follows:

$$S_{X,Y} = 0.50^d \quad \Rightarrow \quad \log S_{X,Y} = \log 0.50^d = d \log 0.50$$
$$\Rightarrow \quad d = \frac{\log S_{X,Y}}{-0.301}$$

TABLE 2—*Probability of paternity as a function of allele frequencies.*

f_i	P
0.1	1.00000000
0.2	1.00000000
0.3	0.99999996
0.4	0.99998874
0.5	0.99902439
0.6	0.96391203
0.7	0.55033568

TABLE 1—*Sample computation of similarity measure.*

	A		*B*		*C*		*D*	
A	$m^i_u = 2$ $m^i_c = 2$	$S_{A,A} = 1$						
B	$m^i_u = 3$ $m^i_c = 1$	$S_{A,B} = 0.5$	$m^i_u = 2$ $m^i_c = 2$	$S_{B,B} = 1$				
C	$m^i_u = 3$ $m^i_c = 1$	$S_{A,C} = 0.5$	$m^i_u = 4$ $m^i_c = 0$	$S_{B,C} = 0$	$m^i_u = 2$ $m^i_c = 2$	$S_{C,C} = 1$		
D	$m^i_u = 3$ $m^i_c = 1$	$S_{A,D} = 0.5$	$m^i_u = 3$ $m^i_c = 1$	$S_{B,D} = 0.5$	$m^i_u = 3$ $m^i_c = 1$	$S_{C,D} = 0.5$	$m^i_u = 2$ $m^i_c = 2$	$S_{D,D} = 1$

The value of d is prominent in the choice of the number of loci used in DNAc. Because the number of loci used leads to a given similarity measure $S_{X,Y}$, which in turn leads to a given kinship degree d according to the formula above, in DNAc, the number of loci is incremented until the value of d converges to an integer. The initial number of loci used depends upon the number of wells in one electrophoresis apparatus.

Clustering Algorithm

A hierarchical clustering approach outputs a structured set of clusters, which is more informative and useful than the unstructured set of clusters returned by partition-based clustering (7). It allows more accurate detection of complex clusters, which is a major advantage when it comes to determining the hierarchical representation of the hereditary relationships and the degree of kinship between different DNA profiles.

Hierarchical clustering can be either "agglomerative" or "divisive." Hierarchical agglomerative clustering (HAC) iteratively merges small clusters into larger clusters, while hierarchical divisive clustering (HDC) iteratively splits clusters into smaller ones. HAC algorithms make clustering decisions based on local patterns, without taking into account the overall structure of objects. HDC algorithms handle the overall structure of objects when making clustering decisions, which makes HDC more accurate than HAC (7).

The literature also reports the graph-theoretical clustering approach based on the maximal spanning tree (MST). A spanning tree of a connected and undirected graph is an acyclic and connected subgraph, which contains all the vertices and some or all branches of that graph. The MST of a weighted graph is the maximally weighted spanning tree of that graph. Like the HDC approach, clustering algorithms based on MST take advantage of the overall structure of objects in making clustering decisions. They are capable of accurately detecting complex clusters (8).

Almost all of the existing algorithms that have been designed to deal with the challenge of partitioning subsets that include complex clusters are based on a hierarchical approach. They adopt the HAC approach, because HDC is more computationally expensive and conceptually complex. They then decide on a clustering by minimizing the local variance of clusters. For instance, in the DBSCAN (9) and SNN (10) algorithms, dense regions are detected and those that are close are merged; in CURE (11), regions corresponding to the closest pair of representatives are merged; in Chameleon (12) and HBC (13), two clusters are merged if their interconnectivity is comparable to their internal connectivity. However, the major drawback with these approaches is that when they have to decide whether two clusters should be merged, only the variance of the clusters concerned is used, and the global variance of clusters is usually ignored. This local decision may seriously affect the global optimality of the final clustering. In addition, they adopt different strategies, but all based on HAC approach to minimize the local variance, which is known to reduce significantly the chance of reaching a global minimum of variance (7). Moreover, the clustering results depend heavily on user-defined input parameters, for which the tuning usually requires extensive efforts, see Table 3 for details.

The main idea behind our clustering approach is to approximate the overall distribution of objects using an MST and to apply an HDC approach using only the limited inter-object connections in the MST. In fact, a conceptual graph can be built from a set of objects to be clustered, each object being considered as a vertex. From the weighted, complete, and undirected graph modeling the similarities between all possible pairs of objects, the starting point of our approach is to build the spanning tree that spans this graph with the maximum total weight. The clustering is then performed as a hierarchical divisive process. Starting from the global MST, we iteratively select and subdivide the subtree that contains the branch of the current minimal weight into two subtrees by cutting off this branch. Subdivision is repeated until each remaining subtree includes only one vertex. Each subdivision results in an enlarged set of subtrees considered as cluster candidates. This set is evaluated according to a novel criterion that is defined as the ratio between the overall compactness within the candidates and the overall compactness between the candidates. A detailed description is given in the rest of this section.

TABLE 3—*Input parameters of different clustering algorithms.*

Algorithm	Parameter	Description
DBSCAN	Eps	Neighborhood of a point
	MinPts	Minimum number of points
CURE	k	Number of clusters
	α	Shrinking factor
	t	Representative points
Chameleon	k-NN	k-nearest neighbor
	MinSize	Initial clustering
	α	Interconnectivity vs. closeness
HBC	m_{ratio}	Subclusters to be merged
	α	Connectivity
	β	Proximity
SNN	Eps	Density measurement
	MinPts	Choice of core points

Starting from V, the set of vertices representing the set of N objects (i.e., genotypes) to be clustered, we build $T(V,B)$, the MST in which the weight $w(b)$ of a branch $b \in B$ linking two vertices is the similarity between the corresponding objects. Now, let the tree $T_P(V_P,B_P)$ be a subtree of $T(V,B)$, where $V_P \subseteq V$ and $B_P \subseteq B$. We then define the weight $w(T_P)$ of the tree $T_P(V_P,B_P)$ as the average weight of all its branches, as follows:

$$w(T_P) = \frac{\sum_{b \in B_P} w(b)}{|B_P|}$$

The weight $w(T_P)$ captures the compactness information of the group of objects covered by the tree $T_P(V_P,B_P)$. The major advantage of measuring the compactness using MST is its time efficiency and its effectiveness in capturing the compactness of complex clusters with arbitrary structures. For the special case of a subtree with only one vertex, the weight cannot be calculated in the way described earlier. Here, we merely assign to the weight the maximal value "1.0", because an object is considered as a cluster with maximal compactness.

Now, let $b_{\min} \in B_P$ be the branch of minimal weight within the subtree $T_P(V_P,B_P)$, and let $T_L(V_L,B_L)$ and $T_R(V_R,B_R)$ be a bipartition of $T_P(V_P,B_P)$ resulting from cutting off the branch $b_{\min}$. We then define the *cosimilarity* $c(T_P)$ of the tree $T_P(V_P,B_P)$, as follows:

$$c(T_P) = w(b_{\min}) \times \frac{w(T_L) \times w(T_R)}{w(T_L) + w(T_R)}$$

The cosimilarity concept draws its inspiration from Ward's dissimilarity (14), also called *Minimum Variance Clustering*, which has been successfully used for solving HAC problems. The role of Ward and Hook's dissimilarity in HAC is to measure the dissimilarity between clusters to decide whether or not they should be merged (7). The concept of *cosimilarity* proposed here is intended

to measure the compactness of objects covered by the tree $T_P(V_P,B_P)$ considered as a cluster candidate. The concept of *cosimilarity* provides a measure to evaluate whether the subtrees $T_L(V_L,B_L)$ and $T_R(V_R,B_R)$ should belong to the same cluster.

For the special case of a subtree with only one vertex, however, the *cosimilarity* cannot be calculated in the way described earlier, because this tree cannot be subdivided. Here, we assign a predefined value to the *cosimilarity*, which is the only input parameter that needs to be set by the user to tune the clustering result. This has an important impact on the final optimal number of clusters. The optimal number of clusters obtained is inversely proportional to the value of this parameter.

Now, let us consider the set of cluster candidates, any set of subtrees obtained by cutting branches of $T(V,B)$. Hereafter, a subtree in this set is called an *inner-subtree*, and a subtree that does not belong to any *inner-subtree* is called an *outer-subtree*. We use T_{is} to represent the set of *inner-subtrees* and T_{os} to represent the set of *outer-subtrees*. We introduce the concept of *inner-cosimilarity*, C_{ic}, defined as the average *cosimilarity* of *inner-subtrees*, and C_{oc}, the *outer-cosimilarity*, defined as the average *cosimilarity* of *outer-subtrees*, such that:

$$C_{ic} = \frac{1}{|T_{is}|}\sum_{t\subset T_{is}} c(t) \text{ and } C_{os} = \frac{1}{|T_{os}|}\sum_{t\subset T_{os}} c(t)$$

The *inner-cosimilarity* aims to measure the overall compactness within the candidates, while the *outer-cosimilarity* aims to measure the overall compactness between the candidates. Thus, the clustering process will be merely a compromise between maximizing C_{ic} the *cosimilarity* of *inner-subtrees* and minimizing C_{os} the *cosimilarity* of *outer-subtrees*. This can be obtained by the ratio of C_{os} and C_{is}. Therefore, the final clustering choice is obtained by maximizing:

$$\frac{C_{ic}}{C_{os}} = \frac{|T_{os}|\sum_{t\subset T_{is}} c(t)}{|T_{is}|\sum_{t\subset T_{os}} c(t)}$$

However, a naïve approach for making a choice of T_{is} and T_{os} that maximizes the ratio $\frac{C_{oc}}{C_{ic}}$, among all possible choices, has quadratic time complexity. To overcome this drawback, we consider only a small subset of choices that have a good chance of maximizing the ratio $\frac{C_{oc}}{C_{ic}}$. We have adopted a hierarchical method that allows us to make this choice in a linear time. The main idea of this method is that instead of visiting all possible choices of T_{is} and T_{os} belonging to $T(V,B)$, we visit in a top-down way only those trees generated by cutting off the branches of maximum weight, which requires complexity $\Theta(N)$ in time and space.

Results and Discussion

In this paper, we present a new DNA testing method, and to the best of our knowledge is the first specifically developed to retrieve the degree of relatedness among genotypes. Unfortunately, there are no other methods, developed to achieve the same task, to perform comparison experiments. The clustering algorithm presented in our paper is a part of our DNA testing method, and in our experiments, we tested several clustering algorithms, and the obtained results were not so satisfactory for publication; this is mainly caused by the number of objects to be clustered.

To show the effectiveness of DNAc, we performed two blind tests on two groups of individuals. The first test was performed on three members of the same family and an unrelated fourth individual. The second test involved four members from a single family.

TABLE 4—*Similarities given by DNAc for the first family.*

	$\mathcal{A}$	$\mathcal{B}$	$\mathcal{C}$	$\mathcal{D}$
$\mathcal{A}$	1.0000			
$\mathcal{B}$	0.2857	1.0000		
$\mathcal{C}$	0.2387	0.4345	1.0000	
$\mathcal{D}$	0.0383	0.0800	0.0725	1.0000

TABLE 5—*Similarities given by DNAc for the second family.*

	$\acute{\mathcal{A}}$	$\acute{\mathcal{B}}$	$\acute{\mathcal{C}}$	$\acute{\mathcal{D}}$
$\acute{\mathcal{A}}$	1.0000			
$\acute{\mathcal{B}}$	0.4631	1.0000		
$\acute{\mathcal{C}}$	0.1578	0.4629	1.0000	
$\acute{\mathcal{D}}$	0.4210	0.2777	0.4354	1.0000

For the first test, we used 112 loci; for the second, only 32. The lists of all alleles and all loci for each individual in each test are available with this paper as supplementary material (Data S1 and S2).

The first test was performed on four individuals, $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$, and $\mathcal{D}$, where individual $\mathcal{A}$ is the half-brother of $\mathcal{B}$ and $\mathcal{C}$, individual $\mathcal{B}$ is the sister of $\mathcal{C}$, and individual $\mathcal{D}$ is unrelated to any of $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}$. The genetic inheritance results of our new similarity measure are shown in Table 4. Using the previously cited formula for computing d, the degree of kinship, we conclude that $\mathcal{C}$ and $\mathcal{B}$ have a kinship of degree $d = 1$, which means that $\mathcal{C}$ and $\mathcal{B}$ are brother and sister, while $\mathcal{A}$ has a relationship of degree $d = 2$ with both $\mathcal{C}$ and $\mathcal{B}$, which means that $\mathcal{A}$ is their half-brother. The very low similarity of $\mathcal{D}$, the individual unrelated to the family, is in line with the known relationships. The similarity measures obtained show the effectiveness of our new kinship similarity measure in predicting the genetic inheritance of the different individuals. Furthermore, the clustering algorithm was able to group individuals $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}$ in the same cluster, while individual $\mathcal{D}$ was grouped in his own cluster, which also confirms the effectiveness of our new clustering algorithm for grouping even a small number of DNA profiles in the right clusters.

The second test was performed on four individuals $\acute{\mathcal{A}}$, $\acute{\mathcal{B}}$, $\acute{\mathcal{C}}$, and $\acute{\mathcal{D}}$. Individual $\acute{\mathcal{A}}$ is the wife of $\acute{\mathcal{C}}$, while $\acute{\mathcal{B}}$ is their child and $\acute{\mathcal{D}}$ is the child of $\acute{\mathcal{C}}$ and half-sister of $\acute{\mathcal{B}}$. Our new similarity measure yielded the genetic inheritance results shown in Table 5. Using the formula for computing d, we conclude that all the pairs of individuals $\acute{\mathcal{A}}$ and $\acute{\mathcal{B}}$, $\acute{\mathcal{B}}$ and $\acute{\mathcal{C}}$, and $\acute{\mathcal{C}}$ and $\acute{\mathcal{D}}$ have a kinship of degree $d = 1$, while individuals $\acute{\mathcal{B}}$ and $\acute{\mathcal{D}}$ have a kinship of degree $d = 2$. In addition, given the low level of similarity between the pair of individuals $\acute{\mathcal{A}}$ and $\acute{\mathcal{C}}$ and between the pair of individuals $\acute{\mathcal{A}}$ and $\acute{\mathcal{D}}$, we can conclude there is no kinship within the individuals of each pair. These results coincide with the known pairwise relationships of the individuals.

We also performed two clustering tests, the first on the group of all individuals $\acute{\mathcal{A}}$, $\acute{\mathcal{B}}$, $\acute{\mathcal{C}}$, and $\acute{\mathcal{D}}$, and the second on the same group with individual $\acute{\mathcal{B}}$ omitted. In the first test, we obtained one cluster including all the individuals $\acute{\mathcal{A}}$, $\acute{\mathcal{B}}$, $\acute{\mathcal{C}}$, and $\acute{\mathcal{D}}$, while in the second test, we obtained two clusters, the first including individual $\acute{\mathcal{A}}$ and the second including individuals $\acute{\mathcal{C}}$ and $\acute{\mathcal{D}}$.

Because there is no kinship between individual $\acute{\mathcal{A}}$ and individuals, it is normal that in the second test, $\acute{\mathcal{A}}$ should have been

clustered apart from individuals Ć and Ď. In the first test, however, the inclusion of individual B́ in the clustering process shows individual Á as having a kinship with individuals Ć and Ď, a virtual kinship because they were all related to B́. These results once again confirm the effectiveness of our new clustering algorithm in grouping a small number of DNA profiles in suitable clusters.

Conclusion

The likelihood ratio method is applicable when someone wants to confirm or refute a kinship hypothesis, but not for identifying an unknown degree of kinship between different genotypes. This is precisely the reason for proposing DNAc in this paper. DNAc is an alternative to the currently accepted likelihood ratio method in the special case where no kinship hypothesis is available.

In this paper, we have developed DNAc, an effective methodology for rapidly generating DNA profiles and clustering them according to kinship. DNAc circumvents the difficulties presented by existing DNA testing methods, including the need for population allele frequencies; it more accurately and systematically highlights the relations of the clustered DNA profiles; and it is effective even with small data sets. It thus provides laboratories with a new and more useful and attractive instrument for generating and analyzing DNA profiles.

Conflict of interest: The authors have no relevant conflicts of interest to declare.

Acknowledgments

We thank Dr. Joe Clarke and Ms. Rina Kampeas for important comments and contributions to the editing of the manuscript.

References

1. Botstein D, White RL, Skolnick M, Davis RW. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am J Hum Genet 1980;32:314–31.
2. Hammond HA, Jin L, Zhong Y, Caskey CT, Chakraborty R. Evaluation of 13 short tandem repeat loci for use in personal identification applications. Am J Hum Genet 1994;55:175–89.
3. Turnpenny P, Ellard S. Emery's elements of medical genetics, 13th edn. London: Elsevier, Churchill Livingstone, 2007.
4. Gjertson DW, Brenner CH, Baur MP, Carracedo A, Guidet F, Luque JA, et al. ISFG: recommendations on biostatistics in paternity testing. Forensic Sci Int, Genet 2007;1:223–31.
5. Biedermann A, Taroni F, Garbolino P. Equal prior probabilities: can one do any better? Forensic Sci Int 2007;172:85–93.
6. Curran T. Forensic DNA analysis: technology and application, BP 443E. Parliamentary Information and Research Service, Library of Canadian Parliament. Ottawa, Ontario, Canada: Parliament of Canada, Information Service, 2007.
7. Manning CD, Raghavan P, Schütze H. Introduction to information retrieval. Cambridge, UK: University Press, 2008.
8. Grygorash O, Zhou Y, Jorgensen Z. Minimum spanning tree based clustering algorithms. Proceedings of the 18th IEEE International Conference on Tools with Artificial Intelligence; 2006 Nov 13–15; Washington, DC. Piscataway, NJ: Institute of Electrical and Electronics Engineers, 2006;73–81.
9. Ester M, Kriegel HP, Sander J, Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise. In: Simoudis E, Han J, Fayyad U, editors. Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining; 1996 Aug 2–4; Portland, OR. Menlo Park, CA: AAAI Press, 1996;226–31.
10. Ertöz L, Steinbach M, Kumar V. Finding clusters of different sizes, shapes, and densities in noisy, high dimensional data. Proceedings of the 2nd SIAM International Conference on Data Mining; 2003 May 1–3; San Francisco, CA. Philadelphia, PA: Society for Industrial & Applied Mathematics, 2003;47–58.
11. Guha S, Rastogi R, Shim K. CURE: an efficient clustering algorithm for large databases. Proceedings of the ACM SIGMOD International Conference on Management of Data; 1998 June 2–4; Seattle, WA. New York, NY: Association for Computing Machinery Special Interest Group on Management of Data, 1998;73–84.
12. Karypis G, Han EH, Vipin K. Chameleon: hierarchical clustering using dynamic modelling. IEEE Comput 1999;8:68–75.
13. Cherng J, Lo M. A hypergraph based clustering algorithm for spatial data sets. Proceedings of the 2001 IEEE International Conference on Data Mining; 2001 Nov 29–Dec 2; San Jose, CA. Piscataway, NJ: Institute of Electrical and Electronics Engineers, 2001;83–90.
14. Ward JH, Hook ME. Application of a hierarchical grouping procedure to a problem of grouping profiles. Educ Psychol Measur 1963;23:69–82.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Data S1. The first set of microsatellite markers used in the study.

Data S2. The second set of microsatellite markers used in the study.

Please note: Wiley-Blackwell is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

Additional information—reprints not available from author:
Ju Yan, Ph.D.
Division of Genetics
Department of Paediatrics
Faculty of Medicine and Health Sciences
University of Sherbrooke
3001 12th Avenue North
Sherbrooke, QC J1H 5N4, Canada
E-mail: ju.yan@usherbrooke.ca