Mutation Rate Model Used in the DNA VIEW Program

Tomasz Krajka

Department of Production Computerisation and Robotisation, Lublin University of Technology,
20-618 Lublin, Poland; t.krajka@pollub.pl

Received: 14 April 2020; Accepted: 20 May 2020; Published: 22 May 2020

Abstract: The first problem considered in this paper is the problem of correctness of a mutation model used in the DNA VIEW program. To this end, we theoretically predict population allele frequency changes in time according to this and similar models (we determine the limit frequencies of alleles—they are uniformly distributed). Furthermore, we evaluate the speed of the above changes using computer simulation applied to our DNA database. Comparing uniformly distributed allele frequencies with these existing in the population (for example, using entropy), we conclude that this mutation model is not correct. The evolution does not follow this direction (direction of uniformly distributed frequencies). The second problem relates to the determination of the extent to which an incorrect mutation model can disturb DNA VIEW program results. We show that in typical computations (simple paternity testing without maternal mutation) this influence is negligible, but in the case of maternal mutation, this should be taken into account. Furthermore, we show that this model is inconsistent from a theoretical viewpoint. Equivalent methods result in different error levels.

Keywords: fingerprinting DNA; paternity testing; DNA VIEW program; mutation model; the rate of mutation; population; Allele frequencies

1. Introduction

Nowadays, DNA polymorphism is widely used in genetic expertise for fixing biological paternity and consanguinity of people. Results of DNA profiling are based on probabilistic and statistical interpretation. This interpretation leads to obtaining the probability of paternity or another relationship determined by Bayesian analysis (likelihood ratio). This probability depends on three factors:

- (i) a priori probability (usually assumed to be equal to 0.5),
- (ii) allele frequencies (main factor), and
- (iii) mutation rate.

The last factor, i.e., mutation rate, is usually small, but in some situations essential. From a viewpoint of population, mutation is the primary generator of genetic diversity, whereas drift tends to reduce the genetic diversity. Most of the human genome has a very low mutation rate: for a typical nucleotide site $\omega = 2 \times 10^{-8}$ (the probability that a mutation will occur during gene transmission from a parent to a child). Short Tandem Repeat (STR) (also known as a microsatellite) mutation rates are even higher (approximately 1 or 2 mutations per locus per thousand generations). However, the key problem is that statistical analysis of the observed mutation cases is therefore almost impossible, because there is not a sufficient amount of observations (cf. [1,2]). Some observations suggest taking the Stepwise Mutation Model (SMM; cf. [3–5]) in which a mutant allele has either $k - 1$ or $k + 1$ repeats, each with a probability of $\frac{1}{2}$, where k is the current repeat number. The SMM is known to be false because the mutation rate increases with allele length and occasionally two-step mutations occur. In [6–8], other defects of the SMM are shown.

Like the Mendel law, the Hardy–Weinberg hypothesis (sometimes called Hardy–Weinberg model/equilibrium/theorem/state of balance) is one of well known genetic laws. The Hardy–Weinberg model describes the genotype and allele frequencies in a non-evolving population. This model has seven basic assumptions (cf. [9,10]):

- organisms are diploid;
- only sexual reproduction occurs;
- generations are nonoverlapping;
- mating is random;
- population size is infinitely large;
- each sex has the same allele frequencies;
- there is no migration, gene flow, admixture, mutation, and selection.

Despite these assumptions, as we show below, there exist mutation models (Model 4) in which the Hardy–Weinberg hypothesis holds. Given these assumptions, the population's genotypes and allele frequencies remain unchanged over successive generations, and the population is said to be in the Hardy–Weinberg equilibrium (cf. [11]). In practice, if we consider loci with $\{A_1, A_2, \dots, A_k\}$ possible alleles and we observe the frequencies of a pair of alleles (A_i, A_j) in the population $P_{i,j}$, $1 \leq i \leq j \leq k$, then the good approximation (as the right-hand side values are estimated only from our database sample, and, as a consequence, these values in the whole population can slightly differ) of allele frequencies is

$$p_i \approx \sum_{1 \leq j \leq k, j \neq i} P_{i,j} + 2P_{i,i}, \quad 1 \leq i \leq k. \quad (1)$$

The Hardy–Weinberg equilibrium can be stated as

$$P_{i,j} \approx \begin{cases} p_i^2, & \text{if } i = j, \\ 2p_i p_j, & \text{otherwise.} \end{cases} \quad 1 \leq i \leq j \leq k. \quad (2)$$

The above approximation is usually (cf. [12] Chapter 3 and [13] Section 3.2) verified by Pearson's χ^2 goodness-of-fit test (cf. [14,15], here arises the problem of rare alleles whose frequencies must be bound together), Fisher's exact test [16–18], or the likelihood ratio test or permutation test (known as the G-test, cf. [19]). Large deviations from the Hardy–Weinberg equilibrium are a consequence of errors in the STR amplification procedure, whereas the small ones follow from the departure from the above assumptions.

In most genetic laboratories, the interpretation of forensic investigations is obtained as a result of computations of Brenner's DNA VIEW computer program [20,21]. In this paper, we concentrate on the considerations of the consequences of the mutation model used in this program which is a generalization of the SMM model in three aspects:

1. The problem of satisfying the Hardy–Weinberg hypothesis and some mathematical inconsistency as a consequence of the assumed mutation model
2. Paternity testing application. We show that in some cases the assumed different mutation models affect different paternity results (exclusion or acceptance of a putative father).
3. The influence of the assumed mutation model on the long-time simulations of a population's behavior.

The aspect of satisfying the Hardy–Weinberg hypothesis is essentially theoretical, whereas the other two are numeric. The considered models are described in Section 2; the materials and methods are described in Section 3; Sections 4–6 are devoted to Problems 1, 2, and 3, respectively; whereas the last section presents conclusions.

Our considerations are similar to those presented in [22]. The main differences are

- (i) they consider similar but not the same mutation models;
- (ii) they state nonstationarity on the basis of the computation of allele frequencies in the consecutive generations, whereas we give a mathematical proof of this assertion;
- (iii) they state some problems in the case of paternity testing, we evaluate them;
- (iv) we focus on some mathematical inconsistency, which they do not deal with; and
- (v) we consider the behavior of a population under the assumed mutation model in the long distance of sequential generations, while they do not consider this problem.

Thus, our approach is wider and deeper than that taken in [22].

2. Materials and Methods

All computations were performed using large database of allelic DNA frequencies obtained from the Medical University of Lublin. This database contains DNA results of 7076 individuals, including personal information such as names, dates, places of birth etc. From this database we took 4410 unrelated persons and computed the distribution of allele frequencies of this sample (for example Table 1 presents the frequencies for the D5S818 locus). Three-thousand-nine-hundred-and-seventy-three individuals were typed using the classical electrophoresis method and 437 by NGS (next generation sequencing). All individuals were genotyped within 17 STR (short tandem repeat) markers located on the autosomal chromosomes. The techniques of DNA amplification are well described in Butler's book [23] in Section 1.

Table 1. Allele (indicated by the number of gene repetition); n_i frequencies; v_i at the D5S818 locus $N = 17$, $\omega_M = 0.00025$, $\omega_F = 0.0012$, $\omega = 0.0011$ (cf. [24]).

i	n_i	v_i	i	n_i	v_i	i	n_i	v_i
1	9	0.00022	7	15	0.15693	13	20	0.02359
2	10	0.00730	8	16	0.17187	14	21	0.00999
3	11	0.01404	9	17	0.13188	15	22	0.00550
4	12	0.09717	10	17.3	0.00011	16	23	0.00157
5	13	0.09998	11	18	0.08324	17	24	0.00056
6	14	0.15850	12	19	0.03752			

The loci of 3973 individuals were amplified using commercial kits: PowerPlex 16Hs System and PowerPlex 17ESX System (Promega Corporation, Madison, WI, USA). PCR products were separated and detected on the ABI 3130 Genetic Analyzer (Applied Biosystems, Foster City, CA, USA). Data were analyzed with the Gene Mapper ID software (v3.2, Applied Biosystems, Foster City, CA, USA, 2005). Laser-induced fluorescence was used in CE systems (Capillary Electrophoresis) with detection (by sensors) as low as 10^{-18} to 10^{-21} mol. The sensitivity of the techniques is attributed to the high intensity of incident light and the ability to accurately focus the light on the capillary. Multi-color fluorescence detection can be achieved by including multiple dichroic mirrors and bandpass filters to separate the fluorescence emission among multiple detectors (e.g., photomultiplier tubes) or by using a prism or grating to project spectrally resolved fluorescence emission onto a position-sensitive detector (sensor) such as a CCD (Charge Coupled Device) array. CE systems with 4- and 5-color LIF (Laser-Induced Fluorescence) detection systems are used routinely for capillary DNA sequencing and genotyping (DNA fingerprint) applications.

The 437 individuals were profiled by next-generation sequences (NGS). Libraries were prepared using the ForenSeq DNA Signature Prep. Kit (Illumina) according to the manufacturer's protocol. Sequencing was performed on MiSeq FGx using the Miseq FGx Reagent Kit v.3 (600 cycles) and the ForenSeq Universal Analysis Software package. Following cluster generation, clusters were imaged using LED and filter combinations specific to each of four fluorescently labeled dideoxynucleotides. After the imaging of a tile was complete, the flow cell was moved into place to expose the next

tile. The process was repeated for each cycle of sequencing. Following image analysis, the software performed base calling, filtering, and quality scoring.

All amplification results were stored in a database. All computations were performed in the R-3.6.0 language on a RSTUDIO platform (v1.2.1335, RStudio, Boston, MA, USA, 2019).

The starting point for our considerations is the following allele frequencies table (for example, for locus D5S818).

Here, and in what follows, N denotes the number of alleles; n_i denotes the number of gene repetitions; i denotes the number of alleles; and $\omega_M, \omega_F, \omega, v_i$ denote the maternal, paternal, and overall mutation rate and the frequency of i -th allele, respectively. It is worthwhile to remark that n_i is not a number but rather a name (17.3 means, 17 full repetitions of gene, and 3 additional pairs of alkali), but throughout the paper, as in [20,21], we will consider n_i as a number.

3. Mutation Models

Throughout this paper, the notation $(i \rightarrow j)$ means allele i was given by parents and arises as allele j in descendant. Although the maternal and the paternal mutations are generally different, we do not take this aspect into account, therefore we do not specify how the parent gives allele i .

Some assumptions (usually assumed) on equal chance of every allele to be transmitted are fulfilled too. Depending on whether mutation takes place or not, the following formula is used to describe the transmission of alleles,

$$(i \rightarrow j) = \begin{cases} 1 - \omega + \omega\omega_{i,i}, & 1 \leq i = j \leq N, \\ \omega\omega_{i,j}, & 1 \leq i, j \leq N, i \neq j, \end{cases} \quad (3)$$

where $\omega_{i,j}, 1 \leq i, j \leq N$ indicates how frequently i mutates into j , provided that i mutates. The different mutation models are derived from a different evaluation of $\omega_{i,j}, 1 \leq i, j \leq N$. The value $\omega_{i,i}$ means that there was mutation, but the number of gene repetition does not change (it is almost not possible to determine this value from observations). Historically, two models were considered:

Model 1.

$$\omega_{i,j} = \frac{1}{N}. \quad (4)$$

Model 2.

$$\omega_{i,j} = \begin{cases} \frac{1}{N-1}, & 1 \leq i, j \leq N, i \neq j, \\ 0, & 1 \leq i = j \leq N. \end{cases} \quad (5)$$

In the description of the DNA VIEW program (Mutation model implemented in DNA VIEW [20]), one can read the following.

Rule A.

Therefore, as a rule of thumb, I suggest assuming that

50% of all mutations increase by one step

50% decrease by one step

5% increase by two steps

5% decrease by two steps

0.5% increase by three steps

0.5% decrease by three steps

... etc.

Never mind that these numbers add to more than 100%.

This description leads us to the following model of mutation.

Model 3.

$$\omega_{i,j} = \begin{cases} \frac{10^{\lfloor -|n_i - n_j| \rfloor}}{\alpha_i}, & 1 \leq i, j \leq N, i \neq j, \\ 0, & 1 \leq i = j \leq N, \end{cases} \quad (6)$$

where

$$\alpha_i = \sum_{\substack{1 \leq j \leq N \\ j \neq i}} 10^{\lfloor -|n_i - n_j| \rfloor}, \quad 1 \leq i \leq N. \quad (7)$$

Previously, in the case of Variable Number of Tandem Repeat (VNTR) investigations, the most popular model of mutation was

Model 4.

$$\omega_{i,j} = v_j, 1 \leq i, j \leq N. \quad (8)$$

What assumptions should the mutation model satisfy? At first, we remark that a correctly defined mutation model should be such that

$$\forall_{1 \leq i \leq N} \sum_{j=1}^N (i \rightarrow j) = 1, \quad (9)$$

because the i -th allele must be transmitted “to somebody”. On the other hand, the second condition deals with the allele frequencies in parent and child populations. At first, allele j has the frequencies v_j , whereas after transmission, it has the frequency $\sum_{i=1}^N (i \rightarrow j)v_i$. Thus, we have

$$\forall_{1 \leq j \leq N} \sum_{i=1}^N (i \rightarrow j)v_i = v_j. \quad (10)$$

The last condition is a formulation of the Hardy–Weinberg hypothesis described in the Introduction.

Using (3), these conditions can be formulated in an equivalent form:

Condition I.

$$\forall_{1 \leq i \leq N} \sum_{j=1}^N \omega_{i,j} = 1, \quad (11)$$

Condition II.

$$\forall_{1 \leq j \leq N} \sum_{i=1}^N \omega_{i,j}v_i = v_j. \quad (12)$$

Furthermore, it is easy to check that in some models (1 and 4) the computed total mutation is not equal to the observed ω . In Model 1 it is $(1 - \frac{1}{N})\omega$ and for the i -th allele in Model 4 $(1 - v_i)\omega$ (instead of ω in both cases). This is a small disturbance in the observed mutation rate.

4. Hardy–Weinberg Equilibrium

In this section, we consider how the above mutation models satisfy conditions I and II (Hardy–Weinberg hypothesis).

Theorem 1. Let us consider the sequence of positive numbers $\{\alpha_i, 1 \leq i \leq N\}$ and the array of non-negative numbers $\{\beta_{i,j}, 1 \leq i, j \leq N\}$ such that

- (i) $\beta_{i,j} > 0, \beta_{i,i} \geq 0, 1 \leq i \neq j \leq N,$
- (ii) $\beta_{i,j} = \beta_{j,i}, 1 \leq i, j \leq N,$
- (iii) $\alpha_i = \sum_{j=1}^N \beta_{i,j} = \sum_{j=1}^N \beta_{j,i}, 1 \leq i \leq N.$

Let us define a matrix

$$A_N = \begin{bmatrix} -\alpha_1 + \beta_{1,1} & \beta_{1,2} & \beta_{1,3} & \dots & \beta_{1,N} \\ \beta_{2,1} & -\alpha_2 + \beta_{2,2} & \beta_{2,3} & \dots & \beta_{2,N} \\ \beta_{3,1} & \beta_{3,2} & -\alpha_3 + \beta_{3,3} & \dots & \beta_{3,N} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \beta_{N,1} & \beta_{N,2} & \beta_{N,3} & \dots & -\alpha_N + \beta_{N,N} \end{bmatrix}, \quad (13)$$

then $\text{rank}(A_N) = N - 1$.

Proof of Theorem 1. Let $A_k, 1 \leq k \leq N$ denote the k -th main minor of the matrix A_N . Assume contrary that $\det(A_k) = 0$ for some $1 \leq k < N$ (cf. Exercise 11 (e) p. 48 [25]). It means that there exist numbers $\lambda_1, \lambda_2, \dots, \lambda_{k-1}$ such that

$$\sum_{i=1}^{k-1} |\lambda_i| > 0, \quad (14)$$

and

$$v_k = \sum_{i=1}^{k-1} \lambda_i v_i, \quad (15)$$

where v_i denotes the i -th column of the matrix A_k . It can be rewritten (after renaming rows and columns in the minor) precisely as

$$\begin{cases} \beta_{k,1} &= (-\alpha_1 + \beta_{1,1})\lambda_1 + \beta_{1,2}\lambda_2 + \dots + \beta_{1,k-1}\lambda_{k-1}, \\ \beta_{k,2} &= \beta_{2,1}\lambda_1 + (-\alpha_2 + \beta_{2,2})\lambda_2 + \dots + \beta_{2,k-1}\lambda_{k-1}, \\ \vdots & \vdots \\ -\alpha_k + \beta_{k,k} &= \beta_{k,1}\lambda_1 + \beta_{k,2}\lambda_2 + \dots + \beta_{k,k-1}\lambda_{k-1}. \end{cases} \quad (16)$$

Among the numbers $\lambda_1, \lambda_2, \dots, \lambda_{k-1}$ there are positive and non-positive ones. By I_1 we denote a set of positive numbers. This set is nonempty. Indeed, if $I_1 = \emptyset$, then summing up both sides of (16), we get

$$-\alpha_k + \sum_{i=1}^k \beta_{k,i} = \sum_{j=1}^{k-1} (-\alpha_j + \sum_{i=1}^k \beta_{i,j})\lambda_j, \quad (17)$$

and on the left-hand side we have a negative number, whereas on the right hand side it is non-negative.

Let t be an index of the maximal λ_i (if there is more than one such index, we take the smallest one),

$$t = \min\{i : \lambda_i = \max_j \lambda_j\}. \quad (18)$$

Then, taking the t -th equation in (16), we get

$$\beta_{k,t} = (-\alpha_t + \beta_{t,t})\lambda_t + \sum_{i < k, i \neq t} \beta_{t,i}\lambda_i \leq (-\alpha_t + \sum_{i=1}^{k-1} \beta_{t,i})\lambda_t \quad (19)$$

and we have a positive number on the left-hand side of the above inequality and a non-positive number on its right-hand side. Thus, this contradiction leads us to $\det(A_k) \neq 0$, for every $1 \leq k < N$.

Now we prove that $\det(A_N) = 0$. It follows from assumption (iii) which can be rewritten as

$$\begin{cases} -\beta_{N,1} &= -\alpha_1 + \beta_{1,1} + \beta_{1,2} + \dots + \beta_{1,N-1}, \\ -\beta_{N,2} &= \beta_{2,1} + -\alpha_2 + \beta_{2,2} + \dots + \beta_{2,N-1}, \\ \vdots & \vdots \\ \alpha_N - \beta_{N,N} &= \beta_{N,1} + \beta_{N,2} + \dots + \beta_{N,N-1}. \end{cases} \quad (20)$$

□

Corollary 1. (p. 62 in [26] or Definitions 7.1.2 and 7.1.4 and Theorem 7.1.3 in [27]) Let A_N be a matrix satisfying the assumptions of Theorem 1. Let the matrix A'_N be obtained from A_N by multiplying the rows of A_N by positive numbers possibly different for different rows. Then,

$$\text{rank}(A'_N) = N - 1. \quad (21)$$

The same result can be obtained if we multiply columns instead of rows.

Corollary 2. Mutation models 1–3 satisfy Condition I, but Condition II is only satisfied when $\bigwedge_{1 \leq i \leq N} v_i = \frac{1}{N}$. Mutation model 4 always satisfies both conditions.

Proof of Corollary 2. Let us consider the following matrices.

$$A_N^{(1)} = \begin{bmatrix} \frac{1}{N} & \frac{1}{N} & \frac{1}{N} & \cdots & \frac{1}{N} \\ \frac{1}{N} & \frac{1}{N} & \frac{1}{N} & \cdots & \frac{1}{N} \\ \frac{1}{N} & \frac{1}{N} & \frac{1}{N} & \cdots & \frac{1}{N} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{1}{N} & \frac{1}{N} & \frac{1}{N} & \cdots & \frac{1}{N} \end{bmatrix}, \quad A_N^{(2)} = \begin{bmatrix} 0 & \frac{1}{N-1} & \frac{1}{N-1} & \cdots & \frac{1}{N-1} \\ \frac{1}{N-1} & 0 & \frac{1}{N-1} & \cdots & \frac{1}{N-1} \\ \frac{1}{N-1} & \frac{1}{N-1} & 0 & \cdots & \frac{1}{N-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{1}{N-1} & \frac{1}{N-1} & \frac{1}{N-1} & \cdots & 0 \end{bmatrix}, \quad (22)$$

and

$$A_N^{(3)} = \begin{bmatrix} 0 & \frac{10^{\lfloor -|n_1 - n_2| \rfloor}}{\alpha_{n_1}} & \frac{10^{\lfloor -|n_1 - n_3| \rfloor}}{\alpha_{n_1}} & \cdots & \frac{10^{\lfloor -|n_1 - n_N| \rfloor}}{\alpha_{n_1}} \\ \frac{10^{\lfloor -|n_2 - n_1| \rfloor}}{\alpha_{n_2}} & 0 & \frac{10^{\lfloor -|n_2 - n_3| \rfloor}}{\alpha_{n_2}} & \cdots & \frac{10^{\lfloor -|n_2 - n_N| \rfloor}}{\alpha_{n_2}} \\ \frac{10^{\lfloor -|n_3 - n_1| \rfloor}}{\alpha_{n_3}} & \frac{10^{\lfloor -|n_3 - n_2| \rfloor}}{\alpha_{n_3}} & 0 & \cdots & \frac{10^{\lfloor -|n_3 - n_N| \rfloor}}{\alpha_{n_3}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{10^{\lfloor -|n_N - n_1| \rfloor}}{\alpha_{n_N}} & \frac{10^{\lfloor -|n_N - n_2| \rfloor}}{\alpha_{n_N}} & \frac{10^{\lfloor -|n_N - n_3| \rfloor}}{\alpha_{n_N}} & \cdots & 0 \end{bmatrix}. \quad (23)$$

Now, if v_o denotes the vector of allele frequencies of parents' population, then Condition II (Hardy–Weinberg hypothesis) can be rewritten as

$$(A_N^{(k)} - I_N)v_o = 0, \quad k = 1, 2, 3, \quad (24)$$

where I_N denotes the $N \times N$ unit matrix. From Corollary 1 we have $\text{rank}(A_N^{(k)} - I_N) = N - 1, k = 1, 2, 3$. Thus, the subspace of solutions of $A_N^{(k)}x = x, k = 1, 2, 3$ is one-dimensional. On the other hand, it is easy to see that the vector $\bar{v}_N = [\frac{1}{N}, \frac{1}{N}, \dots, \frac{1}{N}]^T$ satisfies $A_N^{(k)}\bar{v}_N = \bar{v}_N, k = 1, 2, 3$, thus it is a unique solution (among the vectors with non-negative coordinates such that $\|x\|_{l_1} = 1$). $\square$

Theorem 2. Let us define the sequence of vectors $\{v_k, k \geq 1\}$ by

$$v_{k+1} = ((1 - \omega)I_N + \omega A_N^{(i)})v_k = T_i v_k, \quad k = 0, 1, 2, \dots, \quad i = 1, 2, 3, \quad (25)$$

where v_o denotes the vector of starting allele frequencies in some loci, N is the number of alleles, and $i = 1, 2, 3$ is the selected model of mutations (v_k denotes the frequencies of alleles in the k -th population applying the i -th model of mutation). Then,

$$\lim_{k \rightarrow \infty} T_i^k v_o = \lim_{k \rightarrow \infty} \underbrace{(T_i \circ T_i \circ \dots \circ T_i)}_{k \text{ times}} v_o = \lim_{k \rightarrow \infty} v_k = \bar{v}_N, \quad i = 1, 2, 3. \quad (26)$$

Proof of Theorem 2. Let us consider the Banach space $\mathfrak{R}^N$ with the norm $\|x\|_{l_1} = \sum_{i=1}^N |x_i|$. Furthermore, let $K = \{x \in \mathfrak{R}^N : \|x\|_{l_1} = 1, 0 \leq x_i \leq 1, i = 1, 2, 3, \dots, N\}$ be a compact, separable subset of the space $\mathfrak{R}^N$. Obviously, the fixed point of the mapping T_i on K is the same as the fixed point of $A_N^{(i)}$

and from Corollary 3 this point is $\bar{v}_N$. The mapping T_i is not a contraction because $\|T_1\| = \|A_N^{(i)}\| = 1$, as the matrix T_i consists of non-negative numbers such that their sums are equal to 1 in each row and each column. The norm is achieved for an arbitrary vector x with $\|x\|_{l_1} = 1$ and all non-negative or all non-positive coordinates. Thus, we cannot apply the Banach Contraction Theorem. However, Theorem 9.4 (and Corollary 9.1) in [28] considers the matrix such as T_i . As a conclusion we get a thesis. $\square$

More information about the determinants and algebra of matrices can be found in [29,30].

5. Mutation Models and Paternity Testing

Let us denote unordered pairs $P = \{p_1, p_2\}$, $M = \{m_1, m_2\}$, $D = \{d_1, d_2\}$ of alleged paternal, maternal, and child's STR profile, respectively, ($V = \{P, D, M\}$). As usual in paternity testing, we consider two hypotheses:

Hypothesis 1 (H_1). *Man P is a father of D assuming that M is a mother of D .*

Hypothesis 2 (H_2). *A man different than P is a father of D assuming that M is a mother of D .*

Now, from the Bayesian formula we get

$$P[H_1|V] = \frac{1}{1 + \frac{P[V|H_2]}{P[V|H_1]} \frac{P[H_2]}{P[H_1]}}, \quad (27)$$

and by putting a priori probabilities equal to 0.5 we obtain

$$P[H_1|V] = \frac{1}{1 + \frac{P[V|H_2]}{P[V|H_1]}}, \quad (28)$$

(for more details of the computations see [10,23,31,32]).

However, the computed probability depends on the mutation model (we will write $P_i[V|H_1]$, $i = 1, 2, 3, 4$). Thus, we compute

$$\rho_{i,j} = |P_i[H_1|V] - P_j[H_1|V]|, 1 \leq i < j \leq 4. \quad (29)$$

The frequency of each result V is a frequency of alleles P and M and from Mendel's law, the probability of achieving from such a pair of parents the child's profile D (it also depends on a mutation, but the differences are negligible). We consider two different situations:

- Maternal mutation situation, i.e., when $\{m_1, m_2\} \cap \{d_1, d_2\} = \emptyset$.
- Without maternal mutation when $\{m_1, m_2\} \cap \{d_1, d_2\} \neq \emptyset$.

We show the maximal values of $\rho_{i,j}$ in Table 2 and the distribution of $\rho_{i,j}$ in Figure 1 (the differences in the case without maternal mutation between all models and in the case when maternal mutation takes place between Models 1 and 2 were lower than 0.02, thus negligible).

Table 2. Maximal difference of probability of paternity values between i -th and j -th model ($\rho_{i,j}$) in two cases: with maternal mutation and without it, at the D5S818 locus.

	Maternal Mutation			No Maternal Mutation		
	Model 2	Model 3	Model 4	Model 2	Model 3	Model 4
Model 1	0.00003	0.9977	0.4237	0.00003	0.0005	0.0003
Model 2		0.9977	0.4237		0.0005	0.0003
Model 3			0.9985			0.0006

Figure 1. Distribution of $\rho_{i,j}$ at locus D5S818 in the case of maternal mutation. Width shows the probability of obtaining $\rho_{i,j} = |P_i[H_1|V] - P_j[H_1|V]|$ value indicated by the height (axis y on a chart).

For example, the maximal value 0.9985 was achieved for $P = (15, 15)$, $D = (7, 15)$, $M = (14, 14)$. In this case, we have the overall mutation rate $\omega = 0.001$, and the frequencies of alleles 7, 14, and 15 are equal to 0.002718, 0.009399, and 0.000566, respectively. Thus, we have the Table 3.

Table 3. Probability that we achieve DNA profiles V , providing man P is a father of D ($P[V|H_1]$), probability that we achieve DNA profiles V , providing P is not a father of D ($P[V|H_2]$), and probability that P is a father of D ($P[H_1|V]$) in the case of $(P, D, M) = ((15, 15), (7, 15), (14, 14))$ at locus D5S818.

Value	Model 1	Model 2	Model 3	Model 4
$P[V H_1]$	2.8308×10^{-15}	3.1450×10^{-15}	1.2801×10^{-20}	7.6925×10^{-17}
$P[V H_2]$	9.3044×10^{-18}	1.0338×10^{-17}	3.4825×10^{-17}	8.7205×10^{-20}
$P[H_1 V]$	0.9967239	0.9967236	0.0003674	0.9988676

The above result follows from the fact that in the case of Model 3, allele 15 is significantly more likely to be a mutation of allele 14 than a transfer from a putative father.

Finally, we verify some remarks on the inconsistency of theoretical equations and their practical application. Let $P(p, s)$ be the probability that two alleles p and s are transferred from the same allele (they are brothers). Then, we have

$$P(p, s) = \begin{cases} (1 - \omega)^2 * v_s, & \text{if } p = s, \\ (1 - (1 - \omega)^2) v_s v_p, & \text{otherwise.} \end{cases} \quad (30)$$

but, on the other hand, we have

$$P(p, s) = \sum_t (t \rightarrow p)(t \rightarrow s) v_t. \quad (31)$$

In the formula (31), the assumed mutation model plays an important role, whereas in (30), only the overall mutation rate ω is important. The maximal differences for locus D5S818 between the values computed using these formulas for different p and s are equal to 0.000172, 0.000164, and 0.000123, and for the same $p = s = 12$, are equal to 0.000197, 0.000271, and 0.000271 for Models 1, 2, and 3, respectively. Model 4 always gives an error of 0. All of them are small, smaller than the mutation rate. Which equation should be chosen by the investigator? With very complicated computations, this difference may become considerable. What about other similar situations?

6. Speed of Convergence of Allele Frequencies

Obviously, the speed of convergence depends on the matrix $A_N^{(i)}$, $i = 1, 2, 3$, the starting point v_0 , and the number of alleles. We measure the speed of convergence by the entropy of frequencies in a population's sequential generations. By entropy we denote

$$H_k^{(i)} = -v_1^{(k)} \log_2 v_1^{(k)} - v_2^{(k)} \log_2 v_2^{(k)} - \dots - v_N^{(k)} \log_2 v_N^{(k)}, \quad (32)$$

where $v_k = [v_1^{(k)}, v_2^{(k)}, v_3^{(k)}, \dots, v_N^{(k)}]^T$ denotes the vector of allele frequencies in the k -th generation of a population submitted to action by the i -th model of mutation. From Theorem 4, we have

$$\lim_{k \rightarrow \infty} H_k^{(i)} = H_\infty = \log_2 N, \quad (33)$$

for $i = 1, 2, 3$. The speed of convergence is described by an approximation in a mean-square sense of behavior of entropy in sequential generations.

Another measure of distance between two probability measures is the so-called Kolmogorov distance. For frequencies of sequential generations, we define the Kolmogorov distance between the k -th generation and the 1st generation:

$$K_k^{(i)} = \max_{1 \leq l \leq N} \left| \sum_{j=1}^l (v_j - v_j^{(k)}) \right|, k \geq 1, \quad (34)$$

and investigate the behavior of the sequence $K_k^{(i)}$.

The third measure of distance of probability distribution is called a similarity index. We investigate

$$I_k^{(i)} = \sum_{j=1}^N \min\{v_j, v_j^{(k)}\}, k \geq 1. \quad (35)$$

The behavior of the investigated values for one typical locus D3S1358 is presented in the following Figure 2.

Figure 2. Investigated values: entropy $H_k^{(i)}$, Kolmogorov distance $K_k^{(i)}$, and similarity index $I_k^{(i)}$ in i -th generation and k -th model, for locus D3S1358.

We only consider Models 1–3, because in Model 4 all values are constant (it satisfies the Hardy–Weinberg hypothesis). We see that Models 1 and 2 (almost identical in the figures) are worse than Model 3, as the entropy of frequencies increases much faster in Model 1 and Model 2. In all models, the speed of convergence decreases with increasing the number of generations. Thus, the increase is the biggest in the first generation.

For all the investigated values $H_k^{(i)}$, $K_k^{(i)}$, and $I_k^{(i)}$ for $i = 1, 2, 3$, the best in a mean-square sense is the logarithmic function:

$$f(k) = a \ln(k) + b, k \in \mathcal{N}. \quad (36)$$

The optimal values for H_k belong to the intervals $a \in (0.097(\text{VWA}), 0.197(\text{FESFPS}))$, $b \in (2.2(\text{TPOX}), 4.156(\text{FESFPS}))$, for K_k to $a \in (0.021(\text{TH01}), 0.062(\text{CSF1PO}))$, $b \in (0.099(\text{VWA}), 0.243(\text{D5S818}))$, and for I_k to $a \in (-0.107(\text{CSF1PO}), -0.037(\text{TH01}))$, $b \in (0.460(\text{FGA}), 0.824(\text{VWA}))$. This means that after about 8000 generations, we obtain almost uniformly distributed allele frequencies.

7. Conclusions

- (i) The mutation model assumed in the DNA VIEW program is not correct. After a long-time simulation (~8000 generations), this model results in almost uniformly distributed frequencies (the value of entropy in the D3S1358 locus case is over 93% its maximal value, much more than the value observed in human population, and the trend is still growing, however much slower than in earlier generations) of alleles but in reality no such uniform distribution at any locus can be observed. The uniformly distributed allele frequencies give the maximal entropy, but in reality, the entropy at different loci is smaller than the maximal one (between 62.9% of the maximal entropy in the case of LPL and 87.9% in the case of Penta E).
- (ii) The speed of convergence $H_k^{(3)}$ is directly proportional to the number of allele N and inversely proportional to the difference between the uniform distribution and the distribution of $\nu^{(o)}$ (expressed by H_1). A statistical analysis gives the formula $H_\infty - H_k^{(3)} = 0.33923 + 0.013062 \times N - 0.138466 \times H_1$ ($p = 0.0146$ for N and $p = 0.00229$ for H_1). In the case of the Kolmogorov distance and indices of similarity, this dependence is not statistically significant.
- (iii) Among the considered models: Models 1, 2, and 3, Model 3 is the best. However, from a viewpoint of the Hardy–Weinberg hypothesis, Model 4 is the best. On the other hand, Models 1 and 4 “change” the observed mutation rate; however, this change is very small. Model 3 is preferred from the viewpoint of genetic observation.
- (iv) The choice of model taken for paternity testing has an essential influence in the case of maternal mutation. The only cause why it is hard to detect is that maternal mutations occur rarely. However, if they do, the differences can be enormous, as it was shown in Table 3 ($P[H_1|V]$ value 0.0003674 in Model 3 suggests that the man almost sure is not a father while values over 0.9967 in other models contradict this statement). In this case, we may get a different conclusion dependent on the taken mutation model (the defendant is accepted or excluded as a father).
- (v) By choosing a model different than Model 4, the researcher can omit the difference between two theoretically identical equations (see (30) and (31)). From a theoretical viewpoint, this is a significant lack of cohesion.

Funding: This research received no external funding.

Acknowledgments: The author would like to express his gratitude to the referees for their constructive comments which led to an improved presentation of the paper.

Conflicts of Interest: The author declares no conflicts of interest.

References

1. Vicard, P.; Dawid, A.P.; Mortera, J.; Lauritzen, S.L. Estimating mutation rates from paternity casework. *Forensic Sci. Int. Genet.* **2008**, *2*, 9–18. [[CrossRef](#)] [[PubMed](#)]
2. Chakraborty, R.; Stivers, D.; Zhong, Y. Estimation of mutation rates from parentage exclusion data: Applications to STR and VNTR loci. *Mutat. Res.* **1996**, *354*, 41–48 [[CrossRef](#)]
3. Kimura, M.; Ohta, T. Stepwise mutation model and distribution of allelic frequencies in a finite population. *Proc. Natl. Acad. Sci. USA* **1978**, *75*, 2868–2872. [[CrossRef](#)] [[PubMed](#)]
4. Valdes, A.M.; Slatkin, M.; Freimer, N.B. Allele frequencies at microsatellite loci: the stepwise mutation model revisited. *Genetics* **1993**, *133*, 737–749. [[PubMed](#)]
5. Caskey, C.T.; Pizzuti, A.; Fu, Y.-H.; Fenwick, R.G., Jr.; Nelson, D.L. Triplet repeat mutations in human disease. *Science* **1992**, *256*, 784–789. [[CrossRef](#)] [[PubMed](#)]

6. Lai, Y.; Sun, F. The relationship between microsatellite slippage mutation rate and the number of repeat units. *Mol. Biol. Evol.* **2003**, *20*, 2123–2131. [[CrossRef](#)] [[PubMed](#)]
7. Calabrese, P.; Durrett, R. Dinucleotide repeats in the Drosophila and human genomes have complex, length-dependent mutation processes. *Mol. Biol. Evol.* **2003**, *20*, 715–725. [[CrossRef](#)] [[PubMed](#)]
8. Whittaker, J.C.; Harbord, R.M.; Boxall, N.; Mackay, I.; Dawson, G.; Sibly, R.M. Likelihood-based estimation of microsatellite mutation rates. *Genetics* **2003**, *164*, 781–787. [[PubMed](#)]
9. Crow, J.F. Hardy, Weinberg and language impediments. *Genetics* **1999**, *152*, 821–825. [[PubMed](#)]
10. Evett, I.W.; Weir, B.S. *Interpreting DNA Evidence: Statistical Genetics for Forensic Scientists*, 1st ed.; Sinauer Associates: Sunderland, MA, USA, 1998.
11. Hartl, D.L.; Jones, E.W. *Genetics: Principles and Analysis*, 4th ed.; Jones and Bartlett Publishers: Burlington, MA, USA, 1998.
12. Weir, B.S. *Genetic Data Analysis II. Methods for Discrete Population Genetic Data*, 2nd ed.; Sinauer Associates: Sunderland, MA, USA, 1996.
13. Buckleton, J.; Triggs, C.M.; Walsh, S.J. *Forensic DNA Evidence Interpretation*, 1st ed.; CRC Press: Boca Raton, FL, USA, 2004.
14. Balakrishnan, N.; Voinov, V.; Nikulin, M.S. *Chi-Squared Goodness of Fit Tests with Applications*; Academic Press: Cambridge, MA, USA, 2013.
15. Lemeshow, S.; Hosmer, D.W., Jr. A review of goodness of fit statistics for use in the development of logistic regression models. *Am. J. Epidemiol.* **1982**, *115*, 92–106. [[CrossRef](#)] [[PubMed](#)]
16. Fisher, R.A. *Statistical Methods for Research Workers*, 5th ed.; Oliver and Boyd: Edinburgh, UK, 1934.
17. Raymond, M.; Rousset, F. An exact test for population differentiation. *Evolution* **1995**, *49*, 1280–1283. [[CrossRef](#)] [[PubMed](#)]
18. Slatkin, M. A correction to the exact test based on the Ewens sampling distribution. *Genet. Res.* **1996**, *68*, 259–260. [[CrossRef](#)] [[PubMed](#)]
19. Guo, S.W.; Thompson, E.A. Performing the exact test of Hardy–Weinberg proportion for multiple alleles. *Biometrics* **1992**, *48*, 361–372. [[CrossRef](#)] [[PubMed](#)]
20. Forensic Mathematics. Available online: <http://dna-view.com/> (accessed on 26 March 2020).
21. Brenner, C. Symbolic kinship program. *Genetics* **1997**, *145*, 535–542. [[PubMed](#)]
22. Dawid, A.P.; Mortera, J.; Pascali, V.L. Non-fatherhood or mutation?: A probabilistic approach to parental exclusion in paternity testing. *Forensic Sci. Int.* **2001**, *124*, 55–61. [[CrossRef](#)]
23. Butler, J.M. *Advanced Topics in Forensic DNA Typing: Interpretation*, 1st ed.; Academic Press: Cambridge, MA, USA, 2014.
24. STRBase SRD 130. Available online: <https://strbase.nist.gov/mutation.htm> (accessed on 26 March 2020).
25. Lütkepohl, H. *Handbook of Matrices*; Wiley: Chichester, UK, 1996.
26. Cohn, P.M. *Elements of Linear Algebra*, 1st ed.; Chapman & Hall/CRC: New York, NY, USA; Washington, DC, USA, 1994.
27. Kuttler, K. *An Introduction to Linear Algebra*; Brigham Young University: Provo, UT, USA, 2007.
28. Goebel, K.; Kirk, W.A. *Topics in Metric Fixed Point Theory*; Cambridge University Press: Cambridge, UK, 1990.
29. Vein, R.; Dale, P. *Determinants and Their Applications in Mathematical Physics*, 1st ed.; Springer: New York, NY, USA, 1999.
30. Muir, T.; Metzler, W.H. *A Treatise on the Theory of Determinants*; Dover Publications Inc.: Mineola, NY, USA, 2003.
31. Aitken, C.G.G. *Statistics and the Evaluation of Evidence for Forensic Scientists*, 1st ed.; John Wiley & Sons Inc.: Hoboken, NJ, USA, 1995.
32. Balding, D.J.; Steele, C.D. *Weight-of-Evidence for Forensic DNA Profiles*, 2nd ed.; John Wiley & Sons: Hoboken, NJ, USA, 2015.