Unique or not ? Fingerprint says **“Yes!”**

参赛队员：刘和松（自动化学院）

石　泉（材料学院）

王希晶（光电学院）

指导教师：李志良

参赛单位：重庆大学

参赛时间：2004 年 2 月 5~9 日

Unique or not ? Fingerprint says **"Yes!"**

Abstract

It is hard to believe that everyone's name is unique to himself. But "are your fingerprints unique to yourself?" many people will answer "yes". In order to prove this problem, we create models to explain it.

At first we create a simple and comprehensible model to quickly and coarsely estimate this problem with fingerprints patterns. It's easy to see that our fingerprints vary in general patterns. There are commonly 9 patterns in our fingerprints. Assume that we can distinguish ***c*** kinds of fingerprints in a certain pattern for sampled population ***A***. With the variety of parameter ***c*** and ***A***, we get an interesting conclusion that even we can distinguish a huge amount of fingerprints in a certain pattern (c=1000 billion) for a local city's population (A=30 million), the probability of two or more people have the same fingerprints, or similar exactly, is nearly 100%. To obtain an available model, we create an advanced model more strictly and precisely. The ridges and furrows in fingerprint have some similar characteristics. We select 3 kinds of minutiae points including their numbers, positions and directions and develop a statistical algorithm to calculate the uniqueness of fingerprints. The answer is that fingerprints are unique, as expected! Those two models are under a primary assumption that all characteristics of a given fingerprint can be well recognized. But, if the fingerprints at hand are mostly incomplete, polluted or distorted, odds of misidentification will occur. Based on our second model, we analyze how the incompleteness and distortion of minutiae points influence precision of fingerprints identification, and obtain some essential requirements for acceptable fingerprints identification.

In order to compare odds of misidentification by fingerprint evidence and that by DNA, we develop a model to analyze the identification by DNA evidence. Restricted by biochemical knowledge, our model is sheerly based on the DNA sequence structure, regardless of those complicated biochemical restrictions and requirements. Assuming that two DAN sequences are identical if specific parts of them are the same, the conclusion more strongly remains that DNA is unique, naturally, and odds of misidentification are extremely small(10^{-6}~10^{-7}) even we check only a tiny part of basepairs on the DNA sequence (about 30,000~44,000). It shows the great reliability of DNA identification.

To analyze the stability, sensitivity, strengths and weaknesses of two types of models, we draw the conclusion that *odds of misidentification by DNA evidences are much smaller than that by fingerprints evidences*. And we point out some recommendations to the practices of the models.

Introduction

We develop some models, a pattern model, a shape model and a practical model, to analyze the uniqueness of fingerprints and odds of misidentification by fingerprint evidence. Through comparison, we also develop a sequence model to analyze odds of misidentification by DNA evidence.

Noticing that there are several patterns of fingerprints, and the ridges and furrows in fingerprints are different from one to another, we use these characteristics to develop our model. Considering the incompleteness of the fingerprints, we analyze the probability of misidentification of fingerprints.

Based on our limited knowledge about DNA and genetics, we create a sequence model to calculate the odds. Ignoring those complicated biochemical restrictions and biomolecular requirements, we establish a pure mathematical model related to our problem.

Considering some fundamental assumptions and requirements of both fingerprint and DNA identifications for their accuracy, we propose our opinions on this identification problem.

1. Background

Fingerprints of humankind are formed in the period of a fetus. We know that they are decided genetically by genes which are inherited from their parents. The ridges and furrows formed the patterns or features of fingerprints, and it is commonplace belief that everyone' s fingerprints are unique. Even twinborn babies, their fingerprints are different. Till now, there isn't any case that two persons, either Chinese or American people, have the same fingerprints exactly all over the world. Actually, nobody has found any related record appeared in any governments in the processes of fingerprint management, identification, check and recognition.

In 1953, Watson and Crick[5] put forward the DNA double helix model based on bases pairs. In 1967, Lin and Chargaff[6] found that in single nucleotide chain, there is an balance base distributions: $A \cong T$ and $G \cong C$. In every chromosome of human, the ratio value of basepairs, both A/T and G/C vary in 0.999~1.001. Now according to achievements of Human Genome Projection (HGP)[8] which has been completed by April 14th 2003, the human DNA sequence contains about 3 billion or in more details, 3164.7 million chemical nucleotide bases (A, C, T, and G). Almost all (99.9%) nucleotide bases are exactly the same in all people. In total human genome, any DNA consequence consists of coding and noncoding regions, while the coding regions are called genes. Genes appear to be concentrated in random areas along the genome, with vast expanses of noncoding DNA between. Because of the polymorphism of DNA and the random arrangement of the bases in DNA sequence, it is not hard to imagine the kinds of DNA should be an astronomical figure or a nanoscopic map. And this gives the fundament of DNA identification.

2. Fundamental Assumptions and Hypotheses

2.1 About Fingerprints

- One's fingerprint is immutable throughout his or her life [1], either the person is young or old, its structural characteristics and even refined features are unchanged.
- All characteristics of a given fingerprint can be fully recognized when it is normally acquired. Here, the trace or mark left by a fingerprint is especially defined as a novel term, called "fingertrace" or "fingermark". Therefore, there exists a prefect correspondence relationship between the fingerprint and its fingermark.

Table 1. List of symbols used and their definitions.

symbol	Definition
n	number of macro-classes for fingerprints
m_i	number of subclasses in macro-classes *i*
N	total number of fingerprint classes
c	number of fingerprints that can be recognized in a certain subclasses
S	number of total quantities for fingerprints
A	Number of the sampled population or people under examination
$p_{\mathrm{mod}1,1}$	the probability of two arbitrary persons have the same fingerprint
$p_{\mathrm{mod}1,2}$	the probability of any same fingerprint in the sampled population
Q	number of Minutiae points
Ω_i	Kinds of the shape that a certain kind of minutiae points which can be recognized
φ_i	Acceptability of one minutia point
G	The number of total grids
λ_i	the expected numerical values of minutiae points
l	Reduction coefficient to represent the ratio of the fractional information gotten accurately and the complete information of a fingerprint
N_{alk}	number of bases that differ in single nucleotide chain
N_A	number of adenine in single nucleotide chain
N_G	number of guanine in single nucleotide chain
N_C	number of cytosine in single nucleotide chain
N_T	number of thymine in single nucleotide chain
S_{all}	total kinds of DNA sequences
N_p	total number of population to be analyzed
P	the probability of the fact that at least two person have the perfectly identical DNA sequence

2.2 About DNA

- In single nucleotide chain, $A=T$, $G=C$.

- There are approximately 3 million basepairs of DNA in human different from one person to another. And the remained basepairs are exactly the same.
- The bases are randomly arranged in DNA sequence.
- The contents of A and T are located from 30% to 70%.
- There is no effect mutually among people.

3. Model 1 Primary match of Fingerprint by Rough Pattern Classification

3.1 Analysis and Calculation

According to the pattern of ridges and furrows in fingerprint, we can coarsely classify it into *n* macro-classes and furthermore divide every macro-class into m_i subclasses. Thus the total number of classes is accounted by the following equation:

$$N = \sum_{i=1}^{n} m_i$$

For example, we have a common classification cases as **Figure 1**:

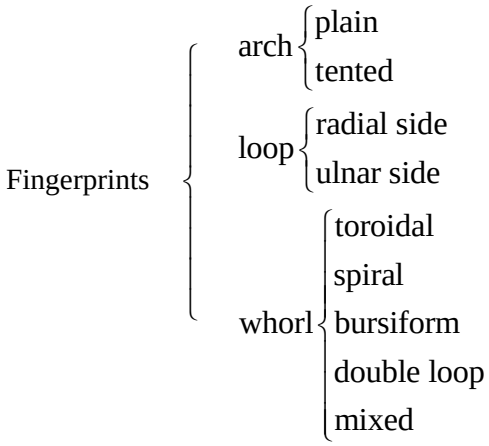

Figure 1. Classification of fingerprints

(1) Consider that we can recognize c kinds of fingerprints in classes j by our eyes and some devices, the total number of fingerprints is $S=cN$, and the probability of any two persons having the same fingerprint is calculated by

$$p_{\text{model1,1}} = 1/S \tag{1}$$

(2) If the number of the sampled population (A) is considered, the probability for the case that at least any two persons have the same fingerprint can be computed as follows:

$$p_{\text{mod1,2}} = 1 - \prod_{i=0}^{A-1} (1 - p_{\text{mod1,1}} i) \tag{2}$$

Noticing that this formula cannot be calculated directly when $p_{\text{mod1,2}}$ is too small. We use another equation to calculate it,

$$p_{\mathrm{mod1,2}} = \frac{p_{\mathrm{mod1,1}} A(A-1)}{2} \tag{3}$$

Through deriving an approximate algorithm, see **Appendix A** for details.

3.2 Results and Discussion

Based on Model 1, one can obtain the results for calculation of probabilities, both the probability, $p_{\mathrm{mod1,1}}$, of two arbitrary persons have the same fingerprint and the probability, $p_{\mathrm{mod1,2}}$, of any same fingerprint in the sampled population, see **Table 2** for details.

Table 2. Probability of the appeared same fingerprints for various classes and populations

N	c	A	S	$p_{\mathrm{mod1,1}}$	$p_{\mathrm{mod1,2}}$
9	10^7	1,000	2.6×10^8	1.1×10^{-8}	0.0055
9	10^8	1,000	9×10^8	1.1×10^{-9}	5.55×10^{-4}
9	10^8	30,000	9×10^8	1.1×10^{-9}	0.393
9	10^8	300,000	9×10^8	1.1×10^{-9}	1.0000
9	10^9	1,000	9×10^9	1.1×10^{-10}	5.55×10^{-5}
9	10^9	30,000	9×10^9	1.1×10^{-10}	0.0488
9	10^9	300,000	9×10^9	1.1×10^{-10}	0.993
10	10^{10}	1,000	1×10^{12}	1.0×10^{-12}	4.99×10^{-7}
10	10^{10}	30,000	1×10^{12}	1.0×10^{-12}	0.00499
10	10^{10}	300,000	1×10^{12}	1.0×10^{-12}	0.0440
10	10^{12}	1,000	1×10^{14}	1.0×10^{-14}	4.995×10^{-9}
10	10^{12}	30,000	1×10^{14}	1.0×10^{-14}	4.50×10^{-6}
10	10^{12}	300,000	1×10^{14}	1.0×10^{-14}	4.50×10^{-4}
10	10^{12}	30,000,000	1×10^{14}	1.0×10^{-14}	0.9889

From **Table 2**, the following phenomena and characteristics can be found:

- The probability of two arbitrary persons having the same fingerprint, $p_{\mathrm{mod1,1}}$, decreases as *c* increases.
- The probability of any same fingerprint in the sampled population, $p_{\mathrm{mod1,2}}$, increases with the increase of A when $p_{\mathrm{mod1,1}}$ is kept constant.
- Theoretically spoken, no matter how $p_{\mathrm{mod1,1}}$ is very small, it is possible to find two or more identical fingerprints if the sampled population *A* **is tendency to infinite. However, when** $p_{\mathrm{mod1,1}}$ is small enough, even for the relatively large sampled population *A* **,** the population of the world A=6.1billion, for instance, the probability of any same fingerprint, $p_{\mathrm{mod1,2}}$, can still be tiny enough for us to consider that the fingerprint identification is evident.

- In more real cases, the sampled population A is usually very small, such as several hundred people, therefore the probability of any same fingerprint, $p_{\mathrm{mod}1,2}$ is still small enough for us to identify whether a person is the targeted one or not through fingerprint evidences with quite high precision.

3.3 Advantage and Weakness

- This model can simply explain the complexity of fingerprints.
- It is too coarse to do some precise analyses on this model. So, a more precise model must be developed.

4. Model 2 Shape match of fingerprint by detailed pattern identification

Based on model 1, a new, more complete model is developed. Under further examination on the ridges and furrows in fingerprint, some similar and even identical characteristics can be found. According to various references [2], there are 6 to 10 kinds of minutiae points. For simplicity, the former case is taken into account, see Figure 2 for details.

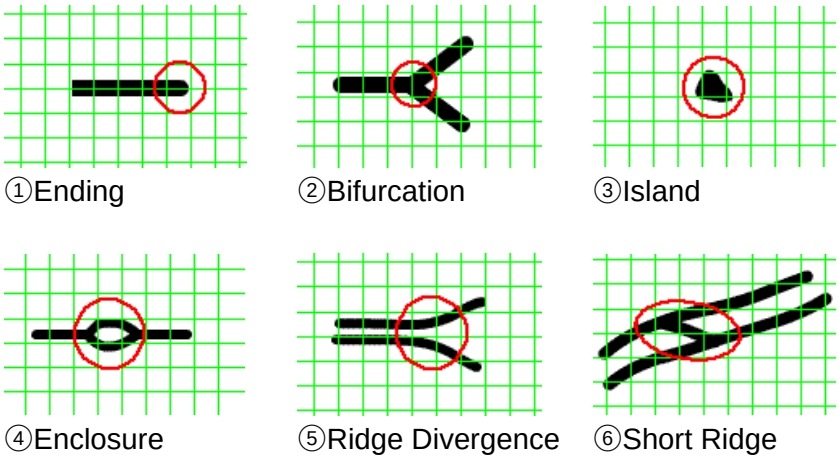

Figure 2. minutiae points in fingerprints

4.1 Additional assumptions

To construct this model, some additional assumptions are required besides basic assumptions, as follows:

- Any single kind of minutiae in one fingerprint conforms to Poisson distribution on its quantities and to uniform distribution on its positions.
- Two fingerprints are regarded to be identical when they have all the same or similarly identical characteristics. Here the word “identical” means the same and/or almost same in the number, the positions and directions of these minutiae points.
- Every fingerprint has its recognition area, and the device employed to identify them has a minimum resolving or called resolution unit as some grids on a fingerprint map. Assume that the recognition area is $2\mathrm{cm}^2$ and the resolution unit is $0.5\mathrm{mm}^2$; the number of total grids count G=20x20=400 and the number of all minutiae points is much smaller than G, see Figure

3 for illustration.

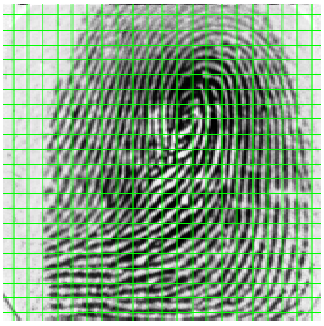

Figure 3. The divided grids on the fingerprint map

- The minimum angel that can be recognized by our eyes or devices is 10^{o} when minutiae points have different directions.
- There never exist two or more minutiae points in a single grid.
- Among those 6 kinds of minutiae points, only 3 kinds are fundamental. They're ①Ending, ② Bifurcation and ③Island as showed in figure 2. Other 3 kinds of minutiae points can be combined by these 3 fundamental kinds. For instance, 2 Bifurcations are properly combined into an Enclosure. Therefore, we only take 3 kinds of fundamental minutiae points into account.

4.2 Model analyses

4.2.1 Numerical distribution of minutiae points

According to the fundamental and additional assumptions and hypotheses, the number of a single kind of minutiae in one fingerprint conforms to Poisson distribution

$$P\{X=k\}=\frac{\lambda^{k}e^{-\lambda}}{k!}$$

So the probability of two fingerprints have the same number of a single kind of minutiae is accounted as:

$$\sum_{k}P\{X=k\}^{2}$$

Then the probability of two fingerprints have the same number of all kinds of minutiae respectively is

$$\prod_{i=1}^{Q}\sum_{k}P_{i}\{X=k\}^{2}$$

4.2.2 Position identification

To simplify this model, we don't consider the trifled differences between the minutiae points of one certain kind, and the total number of minutiae points is much smaller than G, which means every minutiae point can be well recognized. If there are k minutiae points of the same kind, the probability of these minutiae points have same positions in two fingerprints accordingly is

$$p_{G,k}=\frac{1}{C_{G}^{k}}=\frac{k!(G-k)!}{G!}$$

4.2.3 Direction identification

Actually, every minutia point has a direction. Furthermore, we consider the shape feature of the minutiae points (see Table 3 accordingly) and assume that 10^o is the minimum recognized angel.

Table 3. Three fundamental minutiae points: ①Ending, ②Bifurcation and ③Island

Kind of minutiae points	Shape feature	Kinds of shape Ω_i
①Ending	Ending varies in a circle	36
②Bifurcation	Each branch of the Bifurcation takes up 1/3 of the circle	$\frac{12\times12\times12}{3!}=288$
③Island	A single point	1

When simultaneously appears the case that the equalized quantities, the coincided positions and agreed directions, the probability of one given kind of minutiae points in any two fingerprints is estimated by

$$p_G=\frac{\sum_k(p_{G,k}\cdot P\{X=k\}^2)}{\Omega_i}$$

If all kinds of minutiae points are mutually coincided, the probability that two fingerprints are identical is computed as

$$p_{\mathrm{mod}2,1}=\prod_{i=1}^{Q}\frac{\sum_k(p_{G,k}\cdot P\{X=k\}^2)}{\Omega_i} \tag{4}$$

When number of the sampled population A is considered into account， the probability of any same fingerprint in this sampled population is calculated by

$$p_{\mathrm{mod}2,2}=1-\prod_{i=0}^{A-1}(1-p_{\mathrm{mod}2,1}\cdot i) \tag{5}$$

After a statistic of 100 fingerprints, we can easily get the expected numerical values （λ_i） of the 3 fundamental minutiae points, see **Table 4** for more details.

Table 4. The expected numerical values （λ_i） of fundamental minutiae points.

minutiae point	expected numerical value（λ_i）
Ending	24.3
Bifurcation	5.2
Island	0.6

Table 5. Probability of the appeared two identical fingerprints for various classes and populations

G	A	$p_{\mathrm{mod2,1}}$	$p_{\mathrm{mod2,2}}$
300	1,000	9.94×10^{-30}	4.96×10^{-24}
300	300,000	9.94×10^{-30}	4.47×10^{-19}
300	6,100,000,000	9.94×10^{-30}	1.85×10^{-10}
400	1,000	5.91×10^{-30}	2.96×10^{-24}
400	300,000	5.91×10^{-30}	2.66×10^{-19}
400	6,100,000,000	5.91×10^{-30}	1.10×10^{-10}

When different values of both A and G are taken, one can obtain the results for probability of the appeared same fingerprints for various classes and populations see Table 5 for more details.

4.3 Results and Discussion

From the above mentioned results obtained by this model, it is sure that the probability of tow identical fingerprints is extremely small. Even the parameter A is taken as 6.1 billion as the world population, the $p_{\mathrm{mod2,2}}$ is still small enough, at the level of 10^{-10}, so that it is almost impossible to find two persons with the same fingerprint. Therefore, we can naturally believe that the fingerprints are unique for all population in the world.

4.4 Stability analysis: effect of total grids number on stability of model 2

The parameter of grid number G, which represent the resolution of the employed device including our eyes, can vary in a wide range when both resolution area of the examined fingerprint and the resolution power of the applied devices are different. According to Table 6, it is known that the probabilities of the appeared two identical fingerprints is quite stable, in numerical performance for various grid numbers located in a relatively reasonable region, for example, this probabilities $p_{\mathrm{mod2,1}}$ are all kept at the same order of magnitude, the 10^{-30} level, when the grid number is taken from 300 to 5000.

Table 6. Probability of the appeared two identical fingerprints for various classes and populations

G	$p_{\mathrm{mod2,1}}$
300	9.94×10^{-30}
400	5.92×10^{-30}
600	3.53×10^{-30}
1000	2.34×10^{-30}
2000	1.72×10^{-30}
5000	1.43×10^{-30}

4.5 Sensitivity analysis: Effect of the expected numerical values of minutiae

points (λ_i)on stability

When the resolution remains the same (G=400), we change the expected numerical values of minutiae points (λ_i) and examine its effect on stability of model 2. The obtained results can be seen in **Table 7**.

Table 7. Probability of the appeared two identical fingerprints for expected numerical values of minutiae points （λ_i）

λ_1	λ_2	λ_3	$p_{\mathrm{mod2,1}}$
30.0	7.0	0.5	6.21×10^{-36}
28.0	6.0	0.4	1.98×10^{-33}
26.3	5.2	0.3	2.56×10^{-31}
24.3	5.2	0.3	9.94×10^{-30}
22.0	5.0	0.3	1.03×10^{-27}
20.0	4.5	0.3	1.31×10^{-25}
18.0	4.0	0.2	1.57×10^{-23}

From **Table 7**, it is found that:

- $p_{\mathrm{mod2,1}}$ is quite sensitive toward the expected numerical values of minutiae points （λ_i）. While λ_i is the expected value obtained through statistics. If the sampled population *A* is great enough and the statistic technique is quite accurate, it is reasonable to regard these calculated probabilities are reliable.
- The expected numerical values of minutiae points （λ_i） are likely different for different human races and will affect the accuracy of fingerprint evidence. So, the accuracy of fingerprint evidence will in content be dependent on different human races, and besides, the inner relationship seems so complicated for us not to find the regularity,

5. Model 3 Practical match of fingerprint by practical pattern identification

In model 1 and Model 2, which all characteristics of a given fingerprint can be recognized by our eyes or employed devices. Only some ideal situations are considered, such as fingerprints or fingermarks are both complete and clear, which are not the practical ones. But in real cases these obtained fingerprints are mostly incomplete, unclear, polluted, or distorted due to various effecting factors and therefore we can only obtain partial, blurred, contaminated and anamorphic fingerprint samples. Additionally, we will lose some information due to the unreasonable extraction and compression when storing them.

5.1 Analysis and Calculation

To create a more practical match model, based on model 2, we define a reduction coefficient l (l<1) to represent the ratio of the fractional information gotten accurately and the

complete information of a fingerprint. This problem is considered from two factors.

- Partial distortion of fingerprints may occur in the sampling process, if it is influenced by pressure or the other factors. **Figure 4** is the sketch map of two fingerprints from a single thumb.

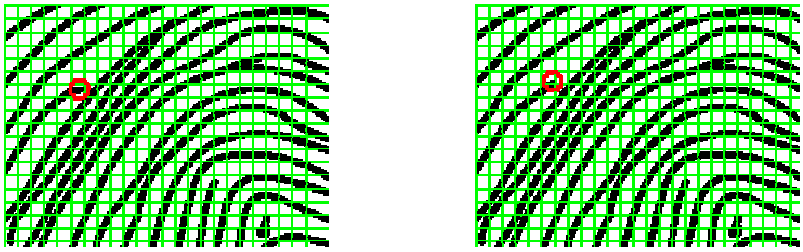

Figure 4. Partial distortion of fingerprints during the sampling process

- The minutiae points are apparently picked up in different grids. If we consider the two minutiae points are different, then misidentification occurs. In order to increase the fault-tolerance of our model, we have to consider and accept this kind of distortion.

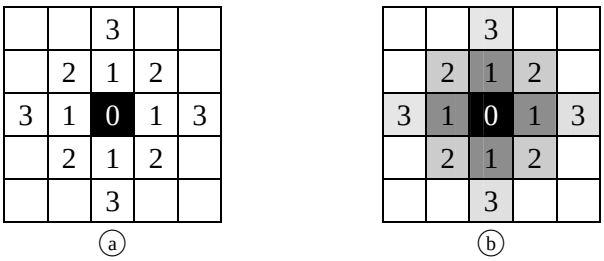

Figure 5. Probability to accept some deviation

If there are one minutia point of fingerprint at grid 0 (Case ⓐ) and another minutia point of fingerprint at grid 0 (Case b), both minutiae points are in same position, then acceptability is 100%. When the minutia points of fingerprint in Case ⓑ are located at grid “1” or “2” or “3”, acceptability are still taken at different degree with the corresponding probabilities being the parameters $\varphi_1, \varphi_2, \cdots$, where one can get:

$$\varphi_i = \begin{cases} 1.00 & i = 0 \\ 0.20 & i = 1 \\ 0.10 & i = 2 \\ 0.01 & i = 3 \\ 0.00 & i \geq 4 \end{cases}$$

Table 8. Probability of the appeared two identical fingerprints
for various Reduction Coefficient *l* together with constant grids and various populations

G	A	l	$p_{\text{mod3,1}}$	$p_{\text{mod3,2}}$
400	1,000	100%	1.41×10^{-27}	7.04×10^{-22}
400	30,000	100%	1.41×10^{-27}	6.35×10^{-19}
400	300,000	100%	1.41×10^{-27}	6.35×10^{-17}
400	1,000,000,000	100%	1.41×10^{-27}	7.05×10^{-10}
400	6,100,000,000	100%	1.41×10^{-27}	2.62×10^{-8}

400	1,000	50%	3.87×10^{-15}	1.93×10^{-9}
400	30,000	50%	3.87×10^{-15}	1.74×10^{-6}
400	3,000,000	50%	3.87×10^{-15}	0.0172
400	34,510,000	50%	3.87×10^{-15}	0.9000
400	6,100,000,000	50%	3.87×10^{-15}	1.0000
400	1000	30%	4.54×10^{-10}	0.0003
400	30,000	30%	4.54×10^{-10}	0.1846
400	100,760	30%	4.54×10^{-10}	0.9000
400	6,100,000,000	30%	4.54×10^{-10}	1.0000
400	1,000	20%	1.63×10^{-7}	0.0781
400	5,319	20%	1.63×10^{-7}	0.9000
400	3,000,000	20%	1.63×10^{-7}	1.0000
400	6,100,000,000	20%	1.63×10^{-7}	1.0000

From **Table 8**, one can find that the probability of the appeared two identical fingerprints is varied in an obvious content with various compression factor l more than with different populations

5.2 Sensitive analyses

To examine the sensitivity of $p_{\text{mod3,1}}$ as l varies, **Figure 6** with a logarithmic scale for $p_{\text{mod3,1}}$ is plotted.

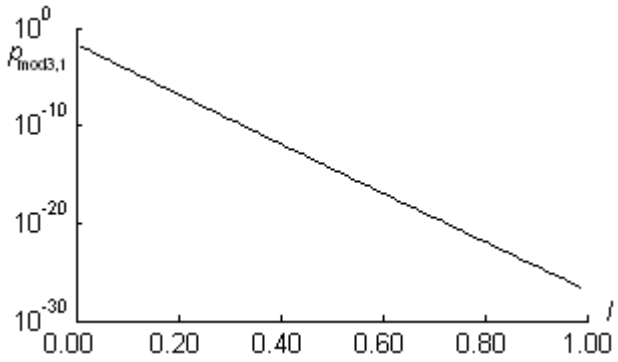

Figure 6. Relationship between probability $p_{\text{mod3,1}}$ and the reduction coefficient l.

Results and Discussion:

- The probability that two arbitrary fingerprints are the same decrease dramatically with the decrease of l, which means $p_{\text{mod3,1}}$ is very sensitive to l.
- we can improve the correctness of identification by enhancing the completeness of fingerprints. On the contrary, if we get only a small part of the fingerprint, the correctness of identification is rather poor.

6. Misidentification of Fingerprint

Fingerprints are frequently used as forensic evidences. Consider a case that we get an incomplete criminal's fingerprints, and we can extract *L??* minutiae points from it . in order

to find out the criminal, we compare the fingerprints with our database which contains A fingerprints samples. And fortunately we find a person's fingerprints have all the L minutiae points same with the criminal's. So, the probability of this person is the criminal, exactly, is [102] (See **Appendix B** for detail)

$$p = 1-(A-1)P(\mathrm{X}) \tag{6}$$

Let the value of L and A varies, we get **Table 9**.

Table 9. Probability p to judge correctly for various L

L	$P(\mathrm{X})$	Sampled population A	Probability p to judge correctly
10	1.10×10^{-11}	100,000	99.9999%
10	1.10×10^{-11}	30,000,000	99.9670%
10	1.10×10^{-11}	6,100,000,000	99.3335%
11	1.73×10^{-12}	30,000,000	99.9948%
11	1.73×10^{-12}	6,100,000,000	99.8946%
12	2.70×10^{-13}	30,000,000	99.9992%
12	2.70×10^{-13}	6,100,000,000	99.9835%
13	4.22×10^{-14}	30,000,000	99.9999%
13	4.22×10^{-14}	6,100,000,000	99.9974%
14	6.58×10^{-15}	6,100,000,000	99.9996%
15	1.03×10^{-15}	6,100,000,000	99.9994%
16	1.60×10^{-16}	6,100,000,000	99.9999%

According to **Table 9**, the probability p to judge correctly is quite large when L is taken 12 and 13 or larger. Now, there are wide applications of fingerprint evidences in many fields, such as the identification recognition, identification validation and so on. In the net bank, electronic business, automated checking system on work attendance, or online trading, etc.

It is worthy to point out that the accuracy of identification by fingerprint evidences is infinitely approached to but not to arrive exactly at 100%, therefore there still exists some erroneous situations. For example, a young, beautiful police woman, Shirley McKie was accused by injustice in 1997 because her fingerprints were analyzed to be the same as that left on a crime scene. Unfortunately and fortunately, it is the only one instance of misidentification until now, which makes the fingerprint evidences are oppugned.

7. Model on DNA

7.1 Problem analysis

Our task is to find out the odds of misidentification of DNA. The factors of misidentification mainly come from both inner (DNA itself) and outer (identify device) causes. Our model is based on the two aspects.

7.2 The polymorphism of DNA Sequences

According to our basic assumption, there're approximately 3 million bases pairs (0.1%)

in human's DNA different from one person to another. So we put the 3 million bases N_{alk} out and make a permutation of them. Consider all four bases A, T, C and G, thus we get:

$$N_{alk} = N_A + N_T + N_C + N_G$$

And the total kinds of DNA sequence made of 3 million different basepairs, S_{all}, is

$$S_{all} = \frac{N_{alk}!}{N_A!N_T!N_C!N_G!}$$

For connivance of calculation, a approximate formula is derived to calculate S_{all} (see **Appendix C** for details)

$$S_{all} = 10^{\left[\left(N_{alk}\ln N_{alk} + \ln\sqrt{2\pi N_{alk}}\right) - \sum_{i=A,G,C,T}\left(N_i \ln N_i + \ln\sqrt{2\pi N_i}\right)\right]\lg e} \tag{7}$$

In computation of S_{all}, only the case that the all four bases in the DNA sequence are arranged as the ratio with $A = T$ and $G = C$. In this process, there are many unreasonable situations, such as one base appears much heavier than the others in a long segment. According to the recent achievement of Human Genome Project and the basic knowledge of chemical biology and molecular genetics, some additional assumptions are made as follows in order to avoid such cases occur and impress some impossible alignments: 1) In some local regions, both A=T and G=C are also arranged. 2) All 3 million different bases are roughly uniformly or homogenously in the total complicate DNA sequence. For various contents of A=T and G=C and for different numbers of segments from the "tandem" ordered DNA sequence of the total human genome, the calculation results are shown in **Table 10**.

Table 10. The calculation results for various basepair contents and different segment numbers

AT\DS	1	23	100	500	1000	5000	10000
0.30	$10^{1.6990\times10^6}$	$10^{1.6989\times10^6}$	$10^{1.6983\times10^6}$	$10^{1.6962\times10^6}$	$10^{1.6938\times10^6}$	$10^{1.6786\times10^6}$	$10^{1.6626\times10^6}$
0.34	$10^{1.7383\times10^6}$	$10^{1.7382\times10^6}$	$10^{1.7376\times10^6}$	$10^{1.7355\times10^6}$	$10^{1.7331\times10^6}$	$10^{1.7177\times10^6}$	$10^{1.7017\times10^6}$
0.38	$10^{1.7683\times10^6}$	$10^{1.7680\times10^6}$	$10^{1.7676\times10^6}$	$10^{1.7655\times10^6}$	$10^{1.6989\times10^6}$	$10^{1.7476\times10^6}$	$10^{1.7315\times10^6}$
0.42	$10^{1.7894\times10^6}$	$10^{1.7893\times10^6}$	$10^{1.7888\times10^6}$	$10^{1.7866\times10^6}$	$10^{1.6989\times10^6}$	$10^{1.7687\times10^6}$	$10^{1.7525\times10^6}$
0.46	$10^{1.8020\times10^6}$	$10^{1.8019\times10^6}$	$10^{1.8013\times10^6}$	$10^{1.7992\times10^6}$	$10^{1.7968\times10^6}$	$10^{1.7812\times10^6}$	$10^{1.7649\times10^6}$
0.50	$10^{1.8062\times10^6}$	$10^{1.8059\times10^6}$	$10^{1.8055\times10^6}$	$10^{1.8033\times10^6}$	$10^{1.8010\times10^6}$	$10^{1.7854\times10^6}$	$10^{1.7691\times10^6}$
0.54	$10^{1.8020\times10^6}$	$10^{1.8019\times10^6}$	$10^{1.8013\times10^6}$	$10^{1.7992\times10^6}$	$10^{1.7968\times10^6}$	$10^{1.7812\times10^6}$	$10^{1.7649\times10^6}$
0.58	$10^{1.7894\times10^6}$	$10^{1.7893\times10^6}$	$10^{1.7888\times10^6}$	$10^{1.7866\times10^6}$	$10^{1.6989\times10^6}$	$10^{1.7687\times10^6}$	$10^{1.7525\times10^6}$
0.62	$10^{1.7683\times10^6}$	$10^{1.7680\times10^6}$	$10^{1.7676\times10^6}$	$10^{1.7655\times10^6}$	$10^{1.6989\times10^6}$	$10^{1.7476\times10^6}$	$10^{1.7315\times10^6}$
0.66	$10^{1.7383\times10^6}$	$10^{1.7382\times10^6}$	$10^{1.7376\times10^6}$	$10^{1.7355\times10^6}$	$10^{1.7331\times10^6}$	$10^{1.7177\times10^6}$	$10^{1.7017\times10^6}$
0.70	$10^{1.6990\times10^6}$	$10^{1.6989\times10^6}$	$10^{1.6983\times10^6}$	$10^{1.6962\times10^6}$	$10^{1.6938\times10^6}$	$10^{1.6786\times10^6}$	$10^{1.6626\times10^6}$

From **Table10**, it is found that: 1) The more the divided segment, less the total kinds of

DNA sequences; however the fluctuations in quantity is small related to itself. 2) The number of total kinds of DNA sequences are still astronomical figures, although only 1% difference come from difference of people. 3) When all four bases are uniformly distributed, as each 25%, the total kinds of DNA sequences are the largest. 4) There exists obvious variation due to various base content, but this variation is quite slight related to itself. So, change of base contents cannot totally convert the fact that huge number of total kinds of DNA sequences is extremely great. From above-mentioned results, it is fully reasonable to believe that different people should perfectly have different DNA sequence characteristics. In the other words, it is completely to distinguish every person, either man or woman, in the whole world, and there exists no erroneous judgment to happen.

Now, knowledge of probability theory is briefly utilized to demonstrate this situation. Actually, among N_p people, the probability of the fact that at least two person have the perfectly identical DNA sequence, *P*, can be calculated according to:

$$P = 1 - \frac{S_{all}(S_{all}-1)\cdots(S_{all}-N_p+1)}{S_{all}^{N_p}}$$

$$= 1 - 10^{\sum_{n=0}^{N_p-1} \lg(S_{all}-n) - N_p \lg S_{all}}$$

For computation, this equation is very complicated and a approximate expression is derived as follows, see Appendix 3-4 for more detailed.

$$P = \frac{N_p(N_p-1)}{2S_{all}} \tag{8}$$

Under the equal amounts of all four bases, A，G，C and T, and proper segmentation of the whole DNA sequence, divided into 10000 segments for instance, the total kinds of DNA sequences can be obtained as: $S_{all} = 10^{1.7691\times10^6}$, see **Table 11**. At the various occasions of N_p, the approximate values of P are computed to be very great.

Table 11 Approximately calculated values of P varied with N_p.

N_p	P
6×10^9	$10^{-1769081.8}$
10×10^8	$10^{-1769082.3}$
3×10^8	$10^{-1769083.4}$
8×10^7	$10^{-1769084.5}$
1×10^7	$10^{-1769086}$
6×10^6	$10^{-1769086.7}$
2×10^6	$10^{-1769087.7}$
2×10^5	$10^{-1769089.7}$
5×10^4	$10^{-1769090.9}$

7.3 Discussion

Based the above-stated results, the probability that two perfectly identical DNA sequence is tendency to zero. On the earth exist nobody that his or her DNA sequence is fully identical to another, although identification is done on all about 6 billion people simultaneously. In fact, the P-value reflects the odds of misidentification. Therefore, we can proudly say, the odds of misidentification by DNA evidences is almost zero. In the other words, it is completely to distinguish every person, either man or woman, in the whole world, and there exists no erroneous judgment to happen; or it is fully reasonable to believe that different people should perfectly have different DNA sequence characteristics.

8. Analysis of the odds of misidentification of DNA

Analysis of the odds of misidentification is established on the basis of all the different bases between human being. They are stochastically distributed in the “tandem” whole DNA sequence consists of 23 chromosomes and only occupied 0.1% of the total around 6 billion basepairs. Obviously, it is very difficult and even impossible to test all bases in the whole DNA sequence for the person. On the other hand, the crime evidence almost provides partial instead of full DNA sequence. So, in practice only special DNA segments are left to be detected.

8.1 Additional Assumptions

- The tested partial segments of 300 different bases stands for the total DNA sequence, and if it is fully identical with the crime sequence and then the whole DNA are also identical.
- There are only 0.1% basepairs that are different due to different people.
- The accuracy of instrumental analysis, such PCR and STR, is very high and its erroneous rates are correspondingly extremely.

8.2 Analysis

Correctly detecting *n* different bases means correspondingly detecting 1000n original bases as the length. Of course, odds of misidentification will occur when all bases cannot be correctly determined. From equation 8, one can easily obtain the probability that the examined person are attributed to the targeted crime due to his or her identical DNA sequence in agreement with the crime’s one for various bases and different population. This probability justly reflects the odds of misidentification by DNA evidence.

Table 12. The probability for both various bases and different population.

N_{alk} \ N_p	6 billion	1.3 billion	200 million	30 million	1 million
3×10^6	$10^{-1806200}$	$10^{-1806200}$	$10^{-1806200}$	$10^{-1806200}$	$10^{-1806200}$
3×10^5	$10^{-180590}$	$10^{-180590}$	$10^{-180590}$	$10^{-1806e00}$	$10^{-180600}$
3×10^4	10^{-18036}	10^{-18037}	10^{-18039}	10^{-18040}	10^{-18043}
3×10^3	$10^{-1781.7}$	10^{-1783}	$10^{-1784.7}$	$10^{-1786.3}$	$10^{-1789.3}$
300	$10^{-157.65}$	$10^{-158.98}$	$10^{-160.61}$	$10^{-162.26}$	$10^{-165.21}$
100	$10^{-37.958}$	$10^{-39.286}$	$10^{-40.912}$	$10^{-42.56}$	$10^{-45.514}$
60	$10^{-14.208}$	$10^{-15.536}$	$10^{-17.162}$	$10^{-18.81}$	$10^{-21.764}$

50	$10^{-8.2894}$	$10^{-9.6179}$	$10^{-11.244}$	$10^{-12.892}$	$10^{-15.846}$
49	$10^{-7.713}$	$10^{-9.0414}$	$10^{-10.667}$	$10^{-12.315}$	$10^{-15.269}$
48	$10^{-7.1286}$	$10^{-8.457}$	$10^{-10.083}$	$10^{-11.731}$	$10^{-14.685}$
47	$10^{-6.5358}$	$10^{-7.8642}$	$10^{-9.4901}$	$10^{-11.138}$	$10^{-14.092}$
46	$10^{-5.9341}$	$10^{-7.2625}$	$10^{-8.8884}$	$10^{-10.536}$	$10^{-13.49}$
45	$10^{-5.3598}$	$10^{-6.6883}$	$10^{-8.3141}$	$10^{-9.9619}$	$10^{-12.916}$
44	$10^{-4.777}$	$10^{-6.1055}$	$10^{-7.7313}$	$10^{-9.3791}$	$10^{-12.333}$
43	$10^{-4.1851}$	$10^{-5.5136}$	$10^{-7.1394}$	$10^{-8.7872}$	$10^{-11.741}$
42	$10^{-3.5835}$	$10^{-4.9119}$	$10^{-6.5378}$	$10^{-8.1856}$	$10^{-11.14}$
41	$10^{-3.0118}$	$10^{-4.3402}$	$10^{-5.9661}$	$10^{-7.6139}$	$10^{-10.568}$
40	$10^{-2.4317}$	$10^{-3.7593}$	$10^{-5.3851}$	$10^{-7.0329}$	$10^{-9.9872}$
39	$10^{-1.8432}$	$10^{-3.1685}$	$10^{-4.7943}$	$10^{-6.4421}$	$10^{-9.3963}$
38	$10^{-1.251}$	$10^{-2.5675}$	$10^{-4.1928}$	$10^{-5.8406}$	$10^{-8.7948}$
37	$10^{-0.71555}$	$10^{-2.0005}$	$10^{-3.6242}$	$10^{-5.272}$	$10^{-8.2262}$
36	$10^{-0.25545}$	$10^{-1.4279}$	$10^{-3.0455}$	$10^{-4.6933}$	$10^{-7.6476}$
35	$10^{-0.019026}$	$10^{-0.86185}$	$10^{-2.4567}$	$10^{-4.1038}$	$10^{-7.058}$
34	$10^{-1.4951e-6}$	$10^{-0.3507}$	$10^{-1.8576}$	$10^{-3.5024}$	$10^{-6.4566}$
33	1	$10^{-0.05278}$	$10^{-1.3009}$	$10^{-2.9377}$	$10^{-5.8919}$
32	1	$10^{-1.2388e-4}$	$10^{-0.75529}$	$10^{-2.3618}$	$10^{-5.3161}$
31	1	$10^{-8.1486e-015}$	$10^{-0.27844}$	$10^{-1.7739}$	$10^{-4.7281}$
30	1	1	$10^{-0.022498}$	$10^{-1.1727}$	$10^{-4.127}$
10	1	1	1	$10^{-0.05287}$	$10^{-1.2228}$
3	1	1	1	$10^{-0.009825}$	$10^{-0.09835}$

From **Table 12**，we can find that it is necessary to increase the length of DNA sequence or segment to decrease the odds of misidentification for the vast of population while it is possible to raise the correctness of identification or to reduce the amount of the involved people for the same tested bases. Therefore, it should be proposed some suggestions to the juristically experts and legal medical experts:

- It is required to extend the tested DNA sequence to enlarge the accuracy of identification.
- It is necessary to ensure the long enough DNA segment to avouch the high precision for vast populations.

8.3 Assessment of Model

Advantages of our model are listed as the follows: 1) It is easily to understand due to the simple and convenient model with effectiveness to describe the diversity of DNA polymorphism by using the knowledge and background of combinatorial mathematics and probabilistic statistics. 2) The results strongly support the positive viewpoints and demonstrate the power of DNA evidence. Both extremely high accuracy and great reliability of DNA evidence illustrate the very low chanciness misidentification. 3) The results with concrete figures and tables provide strong practice and convenience. It is very convenient to obtain directly that only 34 “different” bases, corresponding 34 thousand “original” basepairs, required to be analyzed to identify a population of 1 million, for instance, by DNA evidence with less than one millionth incorrectness. According to these Tables, many practical problems

can conveniently be resolved, such as safety inspection at airports, anti-terrorism progress, business ID check through internet, etc.

8.4 Disadvantage

The main weakness of the model is that there are so many factors to be considered which may affect odds of misidentification. These assumptions and corresponding models should be further established and these works are not yet finished because of the limitation of time and due to the limitation of space. Among them, main factors are included as following:

- Probability of incorrect analysis by the instruments for DNA sequence evidence, especially great and/or branched sequences.
- Effects of relatives. The arrangement of “different” bases in DNA sequence are not fully stochastic but certainly correlated especially for some groups of special species. If suspect has a certain profile, the probability that his relative has the same profile is greater than probability the other profile. But we directly ignored this factor by making assumptions. It would not be reasonable to ignore the effects of relatives when calculating the probability that some person other than a defendant had a particular DNA profile, unless there were good reasons to exclude all relatives. Although we ignored those factors, but we are aware of that, if we had took these factors into consideration, the odds of misidentification by DNA evidence would surely increase.
- According to recent reference[7] it is introduced that only statistics used cannot perfectly address this problem for identification, maybe the mathematical theory of permutation and combination, especially for four bases A, T, G and C, and its combinations with chemical biology and molecular genetics can more powerfully be utilized to overcome our difficulties.

9. Conclusions and Recommendations

According to our models, the odds of misidentification by fingerprints are decided by the sampled population and the number of minutiae points that can be recognized. And the odds of misidentification by DNA are also decided by two factors, the sampled population and the basepairs that can be extracted.

The number of minutiae points on one fingerprint that can be recognized is usually 13 or 14. But it is easy for us to extract more basepairs alone the DNA. So the is nearly zero. And it’s much smaller than the probability of misidentification by fingerprints. We think the odds are about 10^{-8}.

There are also disadvantages of DNA identification. The cost and is high, and the process is complex. If the correctness of fingerprint identification is satisfied, we recommend fingerprints identification. But if the sampled population is big while the minutiae points that can be recognized is not enough, we recommend DNA identification.

10. Our opinions

Both in fingerprint and DNA models created, odds of misidentification are very small. But they are only statistic results based on huge number of cases. It is still a controversial problem that whether it is fit or not to consider a statistic result on a single person. Just as the lottery,

we can easily calculate the probability of winning the prize is $1/10^7$ or $1/10^8$, but there is such a lucky man that he only buys one ticket and wins the top prize. On the contrary, we have a real case of such an unlucky police woman named Shirley McKie convicted of the murder because of the misidentification of fingerprints with an even smaller probability. So we think the usage of fingerprints and DNA identification should be with cautions, especially for forensic evidence just as mentioned by authorized experts. To identify a person only resort to the fingerprints without some other evidences, sometimes is insufficient, or in other words, unfair.

Appendices

Appendix A

The probability of two fingerprints are the same is p_1, and the sample population is A, the probability of two person have the same fingerprints is p_2

Then

$$p_2 = 1 - \prod_{i=0}^{A-1}(1 - p_1 i)$$

When p_2 is too small and cannot directly calculated by computer, use Maclaurin formula

$$\ln(1-x) = -x$$

$$e^{-x} = 1 - x \qquad \text{where x is very small positive number}$$

We can get the approximate value of p_2

Derivation: transform the equation, and the logarithm of the equation is

$$\begin{aligned}\ln(1-p_2) &= \ln\prod_{i=0}^{A-1}(1-p_1 i)\\ &= \sum_{i=0}^{A-1}\ln(1-p_1 i)\\ &= \sum_{i=0}^{A-1}(-p_1 i)\\ &= -p_1\sum_{i=0}^{A-1} i\\ &= -p_1\frac{A(A-1)}{2}\end{aligned}$$

So,

$$\begin{aligned}1-p_2 &= e^{\ln p_2}\\ &= e^{-p_1\frac{A(A-1)}{2}}\\ &= 1-p_1\frac{A(A-1)}{2}\end{aligned}$$

Then

$$p_2 = \frac{p_1 A(A-1)}{2}$$

We use MATLAB to calculate this formula. And we give the script as below:

```
function p2 = p1_to_p2(p1, A)

% if p1 is too small for the computer,
% it cannot be calculated directly,
% and other algorithm should be employed.
if p1 > eps * 10
    temp = 1;
    for i = 0:A-1
        temp = temp * (1 - p1 * i);
    end
    p2 = 1 - temp;
else
    testp = p1 * A * (A - 1) / 2;
    if testp < 10e-3
        p2 = testp;
    else
        p2 = 1 - exp(-testp);
    end
end
```

Appendix B derivation of formula 6

X：an event to coincide the minutiae points with the corresponding site of some fingerprint.

Y：an event to produce this examined fingerprint.

$P(\mathrm{X})$ is the probability of an event X.

$P(\mathrm{Y})$ is the probability of an event Y.

Then the correctly judged probability can be regarded as a conditional probability $P(\mathrm{X}\mid\mathrm{Y})$,

$$P(\mathrm{X}\mid\mathrm{Y}) = \frac{P(\mathrm{XY})}{P(\mathrm{Y})}$$

Because Y is a sub-event of , that is $\mathrm{Y}\subseteq\mathrm{X}$, we have:

$$P(\mathrm{X}\,\mathrm{Y}) = P(\mathrm{X})$$

then

$$P(\mathrm{X}\mid\mathrm{Y}) = \frac{P(\mathrm{X})}{P(\mathrm{Y})}$$

while

$$P(\mathrm{Y}) = 1-[1-P(\mathrm{X})]^{A-1}$$

When $P(\mathrm{X})$ is very small, one can obtain the following equation by using Maclaurin formula:

$$P(\mathrm{Y}) = (A-1)P(\mathrm{X})$$

then

$$P(\mathrm{X}\mid \mathrm{Y}) = \frac{P(\mathrm{X})}{(A-1)P(\mathrm{X})}$$

The correctly judged probability can be obtained as:

$$p = \frac{P(\mathrm{X}\mid \mathrm{Y})}{P(\mathrm{X}\mid \mathrm{Y}) + P(\mathrm{Y})} = \frac{\frac{P(\mathrm{X})}{P(\mathrm{Y})}}{\frac{P(\mathrm{X})}{P(\mathrm{Y})} + P(\mathrm{Y})} = \frac{P(\mathrm{X})}{P(\mathrm{X}) + P(\mathrm{Y})^2} = \frac{1}{1+(A-1)P(X)}$$

In general, the value of $(A-1)P(X)$ is small, then

$$p = 1-(A-1)P(\mathrm{X})$$

In our problem, p is he correctly judged probability and A is the sampled population.

Appendix C

As we know, it is very difficult to calculate *n*! when n is big, so Stirling formula is considered:

$$n! \approx \left(\frac{n}{e}\right)^n \sqrt{2n\pi}$$

We get:

$$\ln n! = n\ln n - n + \ln\sqrt{2n\pi}$$

So,

$$\begin{aligned}
\ln S_{all} &= \ln N_{alk}! - \sum_{i=A,G,C,T} \ln N_i! \\
&= \left(N_{alk}\ln N_{alk} - N_{alk} + \ln\sqrt{2\pi N_{alk}}\right) - \sum_{i=A,G,C,T}\left(N_i\ln N_i - N_i + \ln\sqrt{2\pi N_i}\right) \\
&= \left(N_{alk}\ln N_{alk} + \ln\sqrt{2\pi N_{alk}}\right) - \sum_{i=A,G,C,T}\left(N_i\ln N_i + \ln\sqrt{2\pi N_i}\right)
\end{aligned}$$

$$S_{all} = 10^{\left[\left(N_{alk}\ln N_{alk}+\ln\sqrt{2\pi N_{alk}}\right)-\sum_{i=A,G,C,T}\left(N_i\ln N_i+\ln\sqrt{2\pi N_i}\right)\right]\lg e}$$

Reference

[1]Jin Guangzheng. judicial expertise:p114-134. China politics and law university Express 1995

[2]Geng Xuefeng a discussion of the relationship between the number and characteristics in fingerprintidentification criminal teqnic 1996

[3] ANDREA R. RODDY, MEMBER, IEEE, AND JONATHAN D. STOSZ. Fingerprint

Features—Statistical Analysis and System Performance Estimates

[4] J.L.Blue, G.T.Candela, P.J.Grother, R.Chellappa, and C.L.Wilson, "Evaluation of pattern classifiers for fingerprint and OCR applications," Pattern Recognit., vol.27, no.4, pp.485–501,1994.

[5]Watson J D,Crick F C F.A structure for deoxyribose nucleid acid[J].Nature, 1953,171:737738

[6]Lin H J,Chargaff E. On the denaturation of deoxyribonucleic acid.II. Effcts of concentration[J]. Biochim Biophys Acta,1967,145:398409

[7]ZHONG Dong,ZHAO Gui-jun,ZHANG Zhen-shu,XUN An-long. Advance in the Entire Balance and Local Unbalance of Base Distribution in Genome. Hereditas,2002,(Beijing)24(3): 351355

[8] K. Rao and K. Balck, “Type classification of fingerprints: Asyntactic approach,” IEEE Trans. Pattern Anal. Machine Intell.,vol. PAMI-2, no. 3, pp. 223–231, May 1980.

[9] D.A.Stoney and J.I.Thornton, "A method for the description of minutia pairs in epidermal ridge patterns," J. Forensic Sci., vol.31, no.4, pp.1217–1234, Oct.1986