Journal of Data Science **3**(2005), 295-304

Is the Scientific Discovery of DNA Fingerprint by Chance or by Design?

Harry Yang and Iksung Cho
MedImmune, Inc.

Abstract: DNA fingerprinting is a microbiological technique widely used to find a DNA sequence specific for a microbe. It involves slicing the genomes of the microbe into DNA fragments with manageable sizes, sorting the DNA pieces by length and finally identifying a DNA sequence unique to the microbe, using probe-based assays. This unique DNA is referred to as DNA fingerprint of the microbe under study. In this paper, we introduce a probabilistic model to estimate the chance of identifying the DNA fingerprint from the genome of a microbe when the DNA fingerprinting method is employed. We derive a closed-form functional relationship between the chance of finding the fingerprint and factors that can be experimentally controlled either in part, fully or not at all. Because the odds of finding a specific DNA fingerprint can only be improved by experimental design to a certain degree, in a broader sense, we show that the discovery of a DNA fingerprint is a process governed more by chance than by design. Nevertheless, the results can be potentially used to guide experiments in maximizing the chance of finding a DNA fingerprint of interest.

Key words: DNA fingerprint, partial digestion, PCR, probabilistic model, reproducibility, restriction enzyme.

1. Introduction

In recent years, the application of polymerase chain reaction (PCR) fingerprinting assays has become more common in the accurate and rapid identification of microorganisms (Ben-Ezra, J., Johnson, D.A., Rossi, J., Cook, N., and Wu, A. (1991)). This DNA probe-based technology allows for both discrimination between species and differentiation of isolates belonging to a single species. It is based on either direct amplification of a DNA sequence specific to a microorganism (Belkum, A. (1994)) or generation of an amplified genomic pattern which is highly reproducible (Sobral, B.W.S. and Honeycutt, R.J. (1993)), and which can thus be used as a fingerprint for the species. The development of the former DNA amplification method requires a unique fingerprint of the microorganism.

Several techniques have been developed in the past decade to facilitate the discovery of DNA fingerprints (Belkum, A. (1994)). The methods typically involve slicing large number of copies of species genomes into small pieces using a site-specific enzyme. The DNA fragments are then sorted out according to their base length using gel electrophoresis. Subsequently a few classes will be selected and subjected to PCR amplification using primer specifically designed based on knowledge regarding the species genome. If the PCR method results in replication of a DNA sequence that can be proven to be specific to the genome, the sequence is deemed to be a fingerprint of the microorganism.

The discovery of a DNA fingerprint is a laborious process. It is impacted by experimental conditions such as the efficiency of the restriction enzyme, number of copies of species genomes used in the experiment and lengths of both the genome and DNA fingerprint. In this paper, we introduce a probabilistic model to estimate the chance of identifying a specific DNA sequence of any given length from the genome of a microbe when the DNA fingerprinting method is employed. We establish a functional relationship between the probability of finding a specific DNA sequence, maximum number of fragments into which the DNA sequence can be sliced by restriction enzyme, cutting efficiency of restriction enzyme and number of copies of the microbe genomes used in the fingerprinting experiment. It is shown that the chance of discovering DNA fingerprint can be greatly improved if the enzyme cutting efficiency can be experimentally controlled within a certain range. The model can be potentially used to guide experiments in maximizing the chance of finding a DNA fingerprint of interest. It can also be used to assess the reproducibility of a specific DNA fingerprint discovery. Because the results developed in the paper also imply that the odds of finding of a DNA fingerprint can only be improved by experimental design to a certain degree, in a broader sense, we prove that the discovery of a specific DNA fingerprint of a microbe is governed more by chance than by design.

2. Methods

2.1 Definitions

To facilitate our discussion, we first introduce a few concepts concerning DNA and DNA fingerprinting. DNA is a chemical structure in the chromosomes of living organisms that carries genetic information. It takes the form of a double helix with two strands of genetic material spiraled around each other. Each strand consists a sequence of 4 bases, adenine (A), thymine (T), guanine (G) and cytosine (C), known as nucleotides. The two strands of DNA are chemically bound at each base. The base A will only bond with T, and G with C. In

literature, a DNA sequence is usually described as follows:

$$C\text{-}T\text{-}T\text{-}A\text{-}G\text{-}A\text{-}C\text{-}A\text{-}T\text{-}A\text{-}T$$
$$G\text{-}A\text{-}A\text{-}T\text{-}C\text{-}T\text{-}G\text{-}T\text{-}A\text{-}T\text{-}A$$

DNA strands are read in a particular direction, from the top to the bottom ends. The two ends are referred to as $5'$ (five prime) and $3'$ (3 prime) ends, respectively. To include the directional information of a DNA sequence, the above sequence is often expressed as

$$5'C\text{-}T\text{-}T\text{-}A\text{-}G\text{-}A\text{-}C\text{-}A\text{-}T\text{-}A\text{-}T3'$$
$$3'G\text{-}A\text{-}A\text{-}T\text{-}C\text{-}T\text{-}G\text{-}T\text{-}A\text{-}T\text{-}A5'$$

In this paper, we use the notation $C_1\, C_2\, \ldots\, C_n$ to denote a DNA sequence of n paired bases, with each of the C_i taking the pair of either A-T, G-C, T-A or C-G. The genome of a microbe is the entire DNA sequence in the chromosomes of the microorganism cell that includes all genetic information. The following definitions are also necessary for the development of our method.

Definition 1. (DNA Fingerprint). A sequence of paired nucleotides that is unique to the DNA of a microbe. In this paper, we use Ω to denote a DNA fingerprint.

Definition 2. (Restriction Enzyme). A chemical compound that locates a specific sequence on a DNA and cuts the molecule at that point.

Definition 3. (Restriction Site). A specific sequence on a DNA, at which restriction enzyme cuts the DNA.

Definition 4. (Polymerase Chain Reaction [PCR]). A technique for rapidly multiplying certain segments of DNA; it can produce a million- or billion-fold increase in DNA material within hours.

Definition 5. (Partial Digestion). A collection of DNA fragments which are generated by cutting the DNA sequence of a microbe genome at specific sites, using the restriction enzyme. The cleaving sites, formally called restriction sites, are locations on the DNA where a specific short DNA resides. The word "partial" reflects the fact that a DNA sequence, in a given period of reaction time, might not be completely fragmented at all cutting sites.

Definition 6. (Full Digestion). A collection of DNA fragments of a DNA sequence which is completely fragmented at all restriction sites by a restriction enzyme.

For the rest of the paper, we use the notations Φ, Ω, c and R to denote the entire DNA sequence of a microorganism, DNA fingerprint of the microbe genome, restriction site and enzyme that cuts the restriction site, respectively.

2.2 Modeling of DNA fingerprinting process

The scientific process that leads to the discovery of a DNA fingerprint usually involves the following steps: (1) Isolating the DNA genomes of the microorganism of interest; (2) Cutting the DNA into manageable pieces of different sizes, using restriction enzyme; (3) Sorting the DNA pieces by size. The process by which the size separation, "size fractionation," is done is called gel electrophoresis; (4) Selecting a few sorted DNA pieces, and amplifying the segments, using PCR method, with specially designed primer that binds to a particular sequence of DNA; 5) Amplifying the particular sequence. If this sequence turns out to be specific to the microorganism genome, it can serve as a fingerprint of the microorganism.

In the following, we express the DNA sequence of a microbe as

$$\Phi = B_1 c B_2 c \dots c B_n \tag{2.1}$$

where c is a restriction site on Φ, at which the restriction enzyme R slices Φ. The subsequences $B_i, 1 < i < n$, do not contain c, while B_1 and B_n may contain one c at $5'$ and $3'$ ends, respectively. In a full digestion of Φ, it is cut into n pieces at all cutting sites of c's. Let Ω be the fingerprint of Φ, a sub-string that is unique to Φ. Without loss of generality, we assume that Ω takes the form

$$\Omega = c B_\ell c B_{\ell+1} c \dots c B_{\ell+m-1} c, \tag{2.2}$$

where $\ell > 1$ and $\ell + m - 1 < n$. That is, the fingerprint Ω contains $m + \ell$ restriction sites of c's, with one c being between $B_{\ell-1}$ and B_ℓ, another c between $B_{\ell+m-1}$ and $B_{\ell+m}$. In addition, there are $m - 1$ of the c's in between B_ℓ and $B_{\ell+m-1}$. In a full digestion of the sequence, all c's will be cut, making Ω into m pieces. We refer these c's as $c_0, c_1, \dots, c_m$. Define X_i as random variables that can take value either 0 or 1, with

$$P[X_i = 1] = P[\text{the restriction site } c_i \text{ is cut by the restriction enzyme}] = p.$$

The probability p represents the cutting efficiency of the enzyme. It is reasonable to assume that all X_i are independent. Therefore these $m + 1$ variables X_i are independently identically distributed (iid) according to a Bernoulli distribution. For a partial digestion of Φ to contain the fingerprint Ω, we need

$$X_0 = 1,\ X_1 = X_2 = \dots = X_{m-1} = 0,\ X_m = 1 \tag{2.3}$$

The probability is

$$P[X_0 = 1, X_1 = X_2 = \cdots = X_{m-1} = 0, X_m = 1] = p^2(1-p)^{m-1}. \tag{2.4}$$

2.3 Upper bound on chance of DNA fingerprint discovery

Note that a typical DNA fingerprinting experiment involves many copies, say, r, of the DNA genome under study. Based on the result in (2.4), the following results can be readily verified.

The probability for a DNA probe-based fingerprinting experiment to lead the discovery of a fingerprint is bounded by

$$1 - \left[1 - p^2(1-p)^{m-1}\right]^r \tag{2.5}$$

which achieves its maximum at

$$p = \frac{2}{m+1} \tag{2.6}$$

The number r is the number of copies of the microbe genome used in the experiment, and m is the maximum number of fragments in a full digestion of the fingerprint DNA Ω.

The results in (2.5) and (2.6) suggest that if we could tweak experimental conditions so that the restriction enzyme cutting efficiency can be proportional to the reciprocal of number of fragments in a full digestion of the fingerprint, we could actually maximize our chance for discovering the fingerprint. It is also important to note that the probability in (2.5) is determined not only by the controllable experimental factor r, but also by enzyme efficiency p that can be partially and indirectly manipulated through controlling other experimental factors such as reaction temperature, duration of reaction and etc., and the maximum number of fragments m that a full digestion of the fingerprint Ω possesses. $m+1$ represents the number of restriction sites in the fingerprint (2.2). The factor m, inherent to the microbe DNA, is beyond experimenters' control. Therefore, regardless how well the experiment is designed, the discovery of the DNA fingerprint is always a chance event.

3. Applications

3.1 Reproducibility of DNA Fingerprint

Companies and scientists apply patents for DNA fingerprints they discovered to protect their intellectual rights. Patent application requires the parties to

submit documents detailing experiments that led to the successful findings of the fingerprints. In a recent lawsuit against a company that possesses a patent of the DNA fingerprint of a Microorganism, the patent was argued to be invalid on the ground that five repeat runs, by the plaintiff, of one of the key experiments resulting in a partial digestion of the Microorganism genomes containing the DNA fingerprint did not reproduce the fingerprint. This experiment initially involved the digestion of one billion copies ($r = 10^8$) of the Microorganism genome, using a restriction enzyme. The fingerprint and restriction site consist of 2,500 and 3 pairs of nucleotides, respectively. There are 40 restriction sites residing on the fingerprint, with one at each of the $5'$ and $3'$ ends. In other words, this fingerprint can be fractionated into 39 pieces ($m = 39$) in a full digestion.

In the following, we apply the method developed in the previous section to determining the actual chance of reproducing the DNA fingerprint in five repeat of the original experiment. Let

$$f(p) = 1 - \left[1 - p^2(1-p)^{38}\right]^{10^8} \tag{3.1}$$

By (2.5), $f(p)$ is an upper bound on the probability for a single repeat of the original experiment to contain the DNA fingerprint when the enzyme cutting efficiency is p. A plot of $f(p)$ against p is depicted in Figure 1.

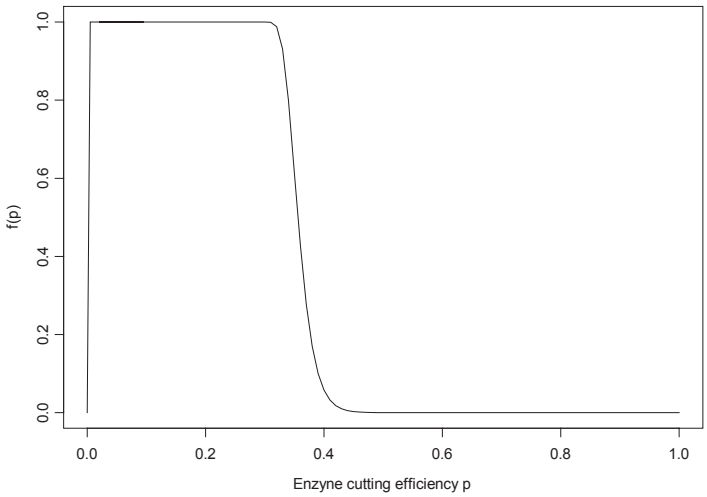

Figure 1: Upper bound on probability of reproducing DNA fingerprint in a single repeat of the original experiment.

As shown in Figure 1, when p determined by the original experimental conditions is no greater than 0.32, $f(p)$ is close to 1. It drops to 0 for $p > 0.4$. For

example, $f(0.32) = 0.99943$ and $f(0.45) = 5.2 \times 10^{-3}$, and $f(0.5) = 9.09 \times 10^{-5}$. If the original experimental condition happens to have resulted in a $p > 0.45$, the chance for the DNA fingerprint to be reproduced in a five repeated experiment is no greater than $5 \times f(0.45) = 0.026$. Reproducing the DNA becomes almost impossible when $p > 0.5$. Pending on p, the DNA fingerprint may or may not reproduced by five repeat runs of the original experiment. Therefore failure of reproducing the DNA fingerprint, in five repeats of the original experiment, cannot be used to discredit the experiment, and invalidate the DNA fingerprint patent.

4. Discussions

In this paper we show that the chance of DNA fingerprint discovery is determined by the number of copies, r, of the microbe genomes, maximum number of DNA fragments m of the fingerprint in a full digestion, and the enzyme efficiency p. If p can be experimentally manipulated to be proportional to the reciprocal of m, by changing controllable experimental factors that impact the enzyme performance, the chance for a successful discovery of the fingerprint can be maximized. This makes intuitive sense because to warrant uncut by the enzyme, the more restriction sites has, the less frequent the enzyme should cleave the original DNA $B_1cB_2 \dots cB_n$ at the restriction sites. Note also that the probability bound $f(p)$ in (2.5) is an increasing function of r. The larger the number of copies r of the microbe genomes used in the experiment is, the higher is the chance of discovery. On the other hand, constricted by resources and experimental conditions, there is a practical limit on the number of copies of microbe genomes that can be used in the experiment. Therefore, to improve one's odds of success, it is helpful to set r at the maximum level that is practically feasible. Another point worthy making is that while the optimal enzyme cutting efficiency $p = 2/(m+1)$ in (2.6) does not depend on r, the chance of success, $f(p)$, can be greatly influenced by r. Figure 2 displays $f(p)$ defined in (3.1) against p for $r = 10^4, 106$ and 10^8. It is evident that for $r = 10^8$, as long as p is no greater than 0.32, the chance of successful identification of the DNA fingerprint is close to one; there is no need to find the optimal $p = 2/(m+1) = 2/(39+1) = 0.05$. For $r = 10^6$, in order to maintain a high probability of success, say, $> 99\%$, p has to be less than 0.23. When only $r = 10^4$ copies of the microbe genomes are used in the experiment, p needs to be in a very close vicinity of the optimal enzyme cutting efficiency, 0.05, to achieve a high success rate.

The results developed in the paper are very useful in guiding experiments intended for discovering DNA fingerprint Ω if one can correlate enzyme efficiency p with the total number of restriction sites of Ω. However, there is usually little prior knowledge about Ω before its discovery. Correlating m, the effect, with p, the cause, is in a way paradoxically, and most definitely not an easy task. In

addition, regardless how well one might design DNA fingerprinting experiments, the chance of success is in part predetermined by factors, such as m, that are out of scientists' control, and others like p that can only be partially and indirectly controlled. Therefore while well-designed experiments can improve one's odds of success in DNA fingerprinting, ultimately it is the inherent properties of a DNA sequence that dictate the chance of success. In other words, the discovery of a DNA fingerprint of a microbe is governed more by chance than by design. Lastly, although the results were derived based on the assumption that the genome of interest possesses a single copy of DNA fingerprint Ω, they can be readily generalized to the case in which the genome has multiple copies of DNA fingerprint.

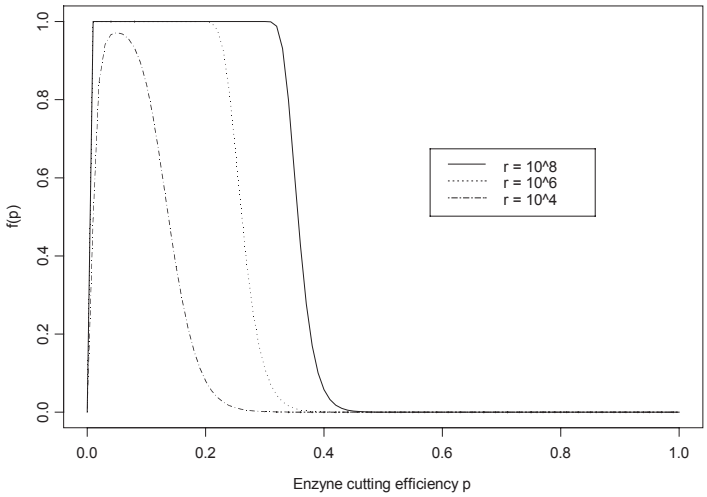

Figure 2: Upper bound probability plots of $f(p)$ for $r = 10^4, 10^6$ and 10^8 .

References

Belkum, A. (1994). DNA fingerprinting of medically important microorganisms by use of PCR. *Clinical Microbiology Reviews* **7**,174-188.

Ben-Ezra, J., Johnson, D. A., Rossi, J., Cook, N., and Wu, A. (1991). Effect of fixation on the amplification of nucleic acids from paraffin-embedded material by the polymerase chain reaction. *J. Histochem. Cytochem.* **39**, 351-354.

Sobral, B. W. S. and Honeycutt, R. J. (1993). High output genetic mapping of polyploids using PCR-generated markers. *Theor Appl Genet.* **86**, 105-112.

Received May 20, 2004; accepted July 4, 2004.

Harry Yang
MedImmune, Inc.
One MedImmune Way
Gaithersbury, MD 20878, USA
yangh@medimmune.com

Iksung Cho
MedImmune, Inc.
One MedImmune Way
Gaithersbury, MD 20878, USA
choi@medimmune.com