Statistical Analysis of STR Data

By Kimberly A. Huston
e-mail genetics@promega.com

Statistical analysis is used to interpret DNA results for genetic identity. In order to determine the significance of a match, it is necessary to support DNA typing results with statistical analysis. These analyses assign a value to the results obtained and enable easier resolution of forensic or paternity cases.

Polymorphic loci contain different sequences at the same locus within and between individuals. These highly variable loci are used in DNA analysis because of their ability to differentiate individuals. Population databases are used to determine the frequency of each allele for a given locus. These databases are generally defined by racial group and geographical region because alleles may have different frequencies in different populations.

Q: What are the Laws of Probability?

A: The First Law of Probability is denoted by the equation:

$$0 \leq \Pr(A/E) \leq 1$$

$$\Pr(A/A) = 1$$

Zero is less than or equal to the probability that A is true, given that E is known, which is less than or equal to one. If we know that A is true, then it has a probability of one.

The Second Law of Probability is denoted by the equation:

$$\Pr(A \text{ or } B/E) = \Pr(A/E) + \Pr(B/E)$$

$$\text{so: } \Pr(\bar{A}/E) = 1 - \Pr(A/E)$$

If A and B are mutually exclusive and E is known, the probability that A or B is true, given E, equals the probability of A, given E, plus the probability of B, given E. Thus, it follows that the probability of A not happening, knowing E, is equal to one minus the probability of A, knowing E.

The Third Law of Probability is denoted by the equation:

$$\Pr(A \text{ and } B/E) = \Pr(A/B, E) \text{ x } \Pr(B/E)$$

The probability of A and B, given that we know E, equals the probability of A occurring, knowing B and E, multiplied by the probability of B occurring, knowing E.

From all of the above laws, the Law of Total Probability follows:

$$\Pr(A) = \Pr(A/B)\Pr(B) + \Pr(A/C)\ \Pr(C)$$

If B and C are two mutually exclusive and exhaustive events, the probability of A is equal to the [probability of A, given B, multiplied by the probability of B] plus [the probability of A given C, multiplied by the probability of C.]

Q: What do exhaustive and exclusive mean?

A: Exhaustive events include all possible outcomes. Exclusive events require that there is no overlap between the outcomes of events.

Q: What does independent mean?

A: Any two events that have no influence on what happens to each other are independent or unassociated. Therefore, for independent events, the probability of both events happening is the product of the probability for each event.

Q: What is meant by being in Hardy-Weinberg Equilibrium?

A: For a population to be in Hardy-Weinberg Equilibrium (HWE), the alleles must be randomly inherited.

Q: What is Bayes' Theorem?

A: Bayes' Theorem is a useful tool for presenting DNA data in a logical manner. The odds form of Bayes' Theorem states:

$$\Pr(E/A) = \frac{\Pr E \text{ and } A}{\Pr(A)} = \frac{\Pr(A/E)\ \Pr(E)}{\Pr(A)}$$

and

$$\Pr(\bar{E}/A) = \frac{\Pr \bar{E} \text{ and } A}{\Pr(A)} = \frac{\Pr(A/\bar{E})\ \Pr(\bar{E})}{\Pr(A)}$$

Taking the ratio of the above equations:

$$\frac{\Pr(E/A)}{\Pr(\bar{E}/A)} = \frac{\Pr(A/E)}{\Pr(A/\bar{E})} \text{ x } \frac{\Pr(E)}{\Pr(\bar{E})}$$

From this, we derive that the posterior odds are equal to the likelihood ratio multiplied by the prior odds.

Prior odds $\left(\frac{\Pr(E)}{\Pr(\bar{E})}\right)$ are odds that are assigned based on initial information.

The likelihood ratio $\left(\frac{\Pr(A/E)}{\Pr(A/\bar{E})}\right)$ is the ratio of two conditional probabilities. By multiplying the prior odds by the likelihood ratio the posterior odds $\left(\frac{\Pr(E/A)}{\Pr(\bar{E}/A)}\right)$ are obtained.

Q: What is conditional probability?

A: All probabilities are conditional based on what we know to be true.

Q: What is the difference between heterozygotes and homozygotes?

A: Heterozygotes are individuals who have two different alleles at the same locus. Homozygotes are individuals who have two identical alleles at a given locus. Heterozygosity is also called the frequency of heterozygotes and is represented by *h* in the following equation.

$$h = \frac{n_h}{n}$$

Where n_h is the number of individual observations with two alleles and n is the total number of individuals.

Since one is either a homozygote or a heterozygote, the frequency of heterozygotes (h) plus the frequency of homozygotes (H) is equal to one.

$$h + H = 1$$

Q: What is matching probability?

A: Matching probability, also known as probability of match (pM), is the number of individuals that may be surveyed before finding the same DNA pattern in a randomly selected individual. This is represented as:

$$pM = \sum_{i=a}^{n} \sum_{j\geq 1}^{n} {P_{ij}}^2$$

Where i and j represent the frequencies of all possible alleles a through n, P_{ij} represents the frequencies of all possible genotypes.

The combined matching probability for more than one locus is the product of the individual matching probability at each locus, assuming that they are not linked.

Q: What does the paternity index represent?

A: The paternity index reflects how many more times likely it is that the person being tested is the biological father, rather than a randomly selected individual. The typical paternity index is assigned to a locus rather than an individual case. Generally, a $PI_{typical}$ of less than one is indicative of non-relatedness. The $PI_{typical}$ is represented by the following equation:

$$PI_{typical} = \frac{1}{2H}$$

The $PI_{typical}$ of several loci is the product of the individual $PI_{typicals}$.

Q: What is the power of discrimination?

A: The power of discrimination is one minus pM. The combined power of discrimination for multiple loci may be calculated by the following equation:

$$P_{d\ combined} = 1 - \prod_{i=1}^{n} (1 - P_{di})$$

Q: What is the Wahlund principle?

A: The Wahlund principle is seen within subpopulations where little gene exchange occurs. In these subpopulations, the allele frequencies differ from those predicted by the Hardy-Weinberg principle. This is seen as an increase in homozygotes and a deficiency in the number of heterozygotes in comparison with the expectations of HWE.

Q: What is the power of exclusion?

A: The power of exclusion, PE, is defined as the fraction of individuals having a DNA profile that is different from that of a randomly selected individual in a typical paternity case. The value for each individual case will vary. The average for a given locus is represented by the following equation.

$$PE = h^2(1–2hH^2)$$

The $PE_{typical}$ for several loci is represented in the following equation:

$$PE_{typical} \overset{n}{\underset{i=1}{=}} 1 - \pi\ (1\text{-}PE_i)$$

Q: What is θ?

A: θ, the co-ancestry coefficient, is the probability that two alleles in the population have descended from the same allele and are identical by descent. This is a measure of the coancestry of populations diverging due to genetic drift. The larger θ, the longer it has been since the populations diverged.

Q: Where can I get more information on statistical calculations?

A: For CODIS calculations, contact the Federal Bureau of Investigation. They can provide Popstats 5.1 software, which will assist in calculations. For advice on paternity case calculations, contact the American Association for Blood Banking at 8101 Glenbrook Rd., Bethesda, MD 20814-2749 or on the Internet at www.aabb.org.

REFERENCES

Brenner, C. and Morris, J. (1990) In. *Proceedings of the International Symposium on Human Identification, 1989,* Promega Corporation.

Evett, I.W. and Weir, B. (1997) *Interpreting DNA Evidence*, Sinauer (in press).