Inferring the Number of Contributors to Mixed DNA Profiles

David R. Paoletti, Dan E. Krane, Michael L. Raymer, and Travis E. Doom

Abstract—Forensic samples containing DNA from two or more individuals can be difficult to interpret. Even ascertaining the number of contributors to the sample can be challenging. These uncertainties can dramatically reduce the statistical weight attached to evidentiary samples. A probabilistic mixture algorithm that takes into account not just the number and magnitude of the alleles at a locus, but also their frequency of occurrence allows the determination of likelihood ratios of different hypotheses concerning the number of contributors to a specific mixture. This probabilistic mixture algorithm can compute the probability of the alleles in a sample being present in a 2-person mixture, 3-person mixture, etc. The ratio of any two of these probabilities then constitutes a likelihood ratio pertaining to the number of contributors to such a mixture.

Index Terms—DNA, mixture, probabilistic computation, optimization, bioinformatics.

1 Introduction

Forensic science has seen the discovery and rapid proliferation of a new tool in the past 20 to 25 years—DNA typing of individuals for the purposes of identification. The use of DNA identification in felony cases has become so common that, as early as 1996, a report prepared for the U.S. Department of Justice noted that "[a]s the use of DNA technology becomes more widely publicized, juries will come to expect it, like fingerprint evidence" [1].

DNA is the genetic material: it is the chemical compound responsible for an organism's ability to inherit traits from its parent(s) [2], [3]. The 23 human chromosomes occur in pairs, one copy from each parent. More than 99 percent of human DNA is identical among even distantly related individuals, but those portions which are different, or polymorphic, can be used for the purpose of human identification. A single region of interest on the chromosome is referred to as a locus, plural loci.

Polymerase chain reaction (PCR) amplification of short tandem repeat (STR) loci have become the method of choice for the purpose of human identification in forensic investigations [4], [5]. Most DNA-typing laboratories use commercially available kits to amplify and label STR alleles associated with evidence and reference samples that are then size fractionated with capillary electrophoresis systems such as the ABI 310 or 3100 Genetic Analyzers [6], [7], or microfabricated chips [8]. Software such as GeneScan and GenoTyper are then used to determine the presence or absence of STR alleles associated with a sample.

In STR loci a short sequence of nucleotides (typically four, or tetranucleotides), is repeated several times in succession. The detectable variability between individuals (polymorphism) is typically based upon differences in the number of times that the tetranucleotide sequences are repeated. Particular repeat counts at a locus give rise to the name of the alleles at the locus and because chromosomes come in pairs, each individual will have two alleles per locus. As an example, Fig. 1 shows a locus, with repeating nucleotide sequence "TCAT," with five repeats on one copy of the chromosome, and seven repeats on the other copy. This individual would be said to have alleles 5 and 7 at this locus, and would be described simply as a "5, 7" for that locus. In some cases, one copy (typically the final copy) of the repeated sequence may not be complete, i.e., "TCA," and is referred to by a floating point number whose integer portion indicates the number of complete repeats and whose decimal portion represents the number of nucleotides left over, so a value such as 9.3 would represent nine complete repeats, with three nucleotides remaining. If an individual exhibits two different alleles at a given locus, such as 5 and 7, the individual is said to be heterozygous at that locus. If the alleles are the same, the individual is said to be homozygous at that locus.

When more than two alleles are observed in the testing results from any single locus, and technical artifacts and noise [9] do not account for any of them, it can be reasonably assumed that the presence of DNA from more than one contributor is the most likely explanation. The absence of a fifth or sixth actual allele is often interpreted as support of there being only two contributors to mixtures even though it is formally possible for the number of contributors to be greater than two [10]. However, if three actual alleles are observed, the sample may arise from a mixture of two individuals, a mixture of three individuals with overlapping alleles, or even a mixture of four or more individuals. Likewise, observing five or more actual alleles

- *D. Paoletti is with the Department of Computer Science, Pennsylvania State University Beaver, 210 Ross Administration Building, 100 University Drive, Monaca, PA 15061. E-mail: drp15@psu.edu.*
- *D. Krane is with the Biological Sciences Department, Wright State University, 128 Biological Sciences Bldg, 3640 Colonel Glenn Hwy, Dayton, OH 45435-0001. E-mail: dan.krane@wright.edu.*
- *M. Raymer and T. Doom are with the Department of Computer Science and Engineering, Wright State University, Dayton, OH, 45435-0001. E-mail: {michael.raymer, travis.doom}@wright.edu.*

Manuscript received 10 Nov. 2008; revised 13 Apr. 2009; accepted 24 May 2010; published online 18 Apr. 2011.
For information on obtaining reprints of this article, please send e-mail to: tcbb@computer.org, and reference IEEECS Log Number TCBB-2008-11-0196.
Digital Object Identifier no. 10.1109/TCBB.2011.76.

1545-5963/12/$31.00 © 2012 IEEE Published by the IEEE CS, CI, and EMB Societies & the ACM

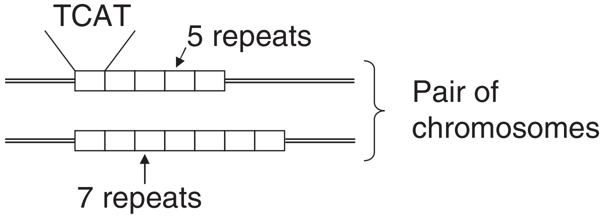


Fig. 1. Hypothetical section of a short tandem repeat, showing the repeated sequences and the repeat counts that give the alleles their designations.

at one locus is an indication of three or more contributors. However, it becomes increasingly difficult to determine the exact number of contributors as the number of observed alleles increases [11].

Counting the maximum number of alleles observed over all loci (and dividing by two) has been a de facto method of determining the minimum number of contributors, which is then used as the actual number of contributors. With this approach, the most likely number of contributors can be determined in up to 75 percent of the cases where the observed number of contributors is less than the actual number. This determination would lead law enforcement personnel to seek out perpetrators that they might not have otherwise known existed.

Lauritzen and Mortera [12] derives an inequality for the probability of observing a particular DNA profile, assuming a fixed number of unknown contributors. $P_x(E|H)$ is the probability that x contributors explain evidence sample E under some hypothesis H:

$$P_x(E|H) \leq \prod_{m=1}^{M} \left(\sum_{a \in E_m} p_a^m \right)^{2x}, \tag{1}$$

where p_a^m denotes the frequency of allele a at locus m. This inequality is then used to obtain a upper bound on the number of unknown contributors who are likely to have created the mixture. Using five loci (LDLR, GYPA, HBGG, D7S8, and Gc), one example demonstration is given using two contributors [12].

There are only three pieces of information that might be helpful in determining the number of contributors to a mixture:

1. The maximum number of alleles at each of the tested loci.
2. Peak height/area, and balance/imbalance.
3. The frequencies of the observed alleles.

A common first step in the analysis of mixed DNA samples is to attempt to determine which alleles in the mixed sample come from each contributor (referred to as mixture deconvolution). While there has been significant work in this area [13], [14], [15], there are still avenues of exploration that can yield useful results. While most existing approaches focus on utilizing the peak height or area information to deconvolve the mixture [16], [17], current research suggests that peak height information will not be helpful in most circumstances, particularly those involving more than two contributors [18]. The alleles themselves, and their frequency in the population, provide additional information which may be helpful in resolving varying numbers of contributors [19], [20].

The approach described in this paper was inspired by the work of [21], which deals with single nucleotide polymorphisms (SNPs). In that formulation there are only two possibilities: either a polymorphism has been observed (with probability p) or it has not (probability $1 - p$). The method presented here has a more complex formulation, due to there being more than two commonly observed alleles per locus. The Probabilistic Mixture Model (PMM) laid out herein will shed light on the number of contributors to a DNA sample by analyzing all possible ways to explain the sample, given different numbers of contributors. For instance, if one has calculated $P(2)$, the sum of the probabilities of all possible ways to explain a sample given two contributors, and $P(3)$, the sum of the probabilities of all explanations for the same sample given three contributors, then the likelihood ratio $LR = \frac{P(2)}{P(3)}$ can be used to determine which explanation is more likely, with all other factors being equal. Counting the number of alleles observed at each locus gives a lower bound on the number of contributors to a mixture, while [12] provides an upper limit. This new approach provides a method of determining the most likely number of contributors.

2 Probabilistic Mixtures

In order to achieve the most accurate misclassification measures, one would need to analyze all possible combinations of all possible profiles. This problem is, to say the least, impractical, there being on the order of $\binom{10^{12}}{3}$ possible combinations for three-person mixtures. However, because the loci are commonly treated as being independent [22], and because of the nature of the final analysis function (the maximum number of alleles observed across all loci in the mixture), it is possible to analyze each locus independently, and combine the results from all loci without resorting to the inclusion-exclusion rule.

These "probabilistic mixtures," make use of the sample allele frequencies from the underlying population, and are only accurate as far as these frequencies are representative of the underlying population [23]. Further, since the computational complexity does not depend on the number of profiles available, but only on the alleles themselves (see Section 2.2), the algorithm scales well: increasing the size of the data set increases accuracy without directly affecting the computational complexity.

2.1 Derivation

Each individual in a mixture contributes two alleles (though they may have the same value, in the case of homozygous individuals). Thus c individuals contribute $2c$ alleles. Referring to the number of unique, visible alleles as u, the remaining, duplicate alleles d are given by $d = 2c - u$. Given that a locus l has $|l|$ commonly observed alleles (in a given population), the probability of observing u unique alleles and d duplicates at locus l is:

$$P(l,c,u) = \underbrace{\sum_{i_1=1}^{|l|-u+1}\sum_{i_2=i_1+1}^{|l|-u+2}\cdots\sum_{i_u=i_{u-1}+1}^{|l|}}_{u}\underbrace{\sum_{j_1=1}^{u}\sum_{j_2=j_1}^{u}\cdots\sum_{j_d=j_{d-1}}^{u}}_{d} P, \tag{2}$$

where P represents the product of the probabilities:

$$P = \underbrace{p_{i_1}p_{i_2}\cdots p_{i_u}}_{u}\underbrace{p_{i_{j_1}}p_{i_{j_2}}p_{i_{j_d}}}_{d}\cdot perms. \tag{3}$$

In the above equation, $perms$ represents the number of permutations of u unique alleles and d duplicates, chosen from the unique alleles. For instance, in a situation with one contributor ($c = 1$), and two unique alleles being chosen ($u = 2$) from locus l, there are zero duplicates ($d = 2c - u = 2 \cdot 1 - 2 = 0$), and

$$P(l,1,2) = \sum_{i_1=1}^{|l|-1}\sum_{i_2=2}^{|l|} p_{i_1}p_{i_2} \cdot perms, \tag{4}$$

where p_{i_1} and p_{i_2} are the allele frequencies for the two unique alleles chosen, and $perms = 2$.

The u unique and d duplicate alleles can be rearranged in $(u + d)!$ ways, but the duplicates produce nonunique permutations. Thus $(u + d)!$ must be divided by the products of the factorials of the numbers of duplicates. A few definitions are needed to describe $perms$. First, define the set of duplicates, D, as:

$$D = \{j_1, j_2, \dots j_d\}. \tag{5}$$

Define a set X to be a subset of D such that it contains all copies of a particular value (i.e., all 2's)

$$X = \{x_i : x_i, x_j \in D, x_i = x_j, i \neq j\}, \quad X \bigcap (D - X) = \emptyset. \tag{6}$$

Now define the number of permutations (perms) as

$$perms = \frac{(u+d)!}{dups}, \tag{7}$$

where $dups$, the correction factor for the nonunique permutations, is

$$dups = \prod_{\forall X \subset D} (|X| + 1)! \tag{8}$$

For example, suppose the locus under consideration has six commonly observed alleles (14, 15, 16, 17, 18, 19), and suppose a three-person mixture ($c = 3$), with four unique alleles and two duplicates (since $d = 2c - u$). Suppose that the unique alleles selected are 14, 15, 16, and 17, and that 14 is chosen to be repeated twice more. Thus $D = \{14, 14\}$, and the only possible $X = \{14, 14\}$. Therefore, $dups = 3! = 6$ (3! since 14 occurs a total of three times, once as a unique, and twice as a duplicate), and $perms = \frac{(u+d)!}{3!} = \frac{6!}{3!} = 120$.

Thus $P(l,\ c,\ u)$ gives, for each locus l, each number of contributors c, and each possible number of visible peaks u (from 1 to $2c$), the probability of observing this situation. The cumulative probability $C(l,\ c,\ u)$ of observing from 1 to u unique alleles at l is therefore:

$$C(l,c,u) = \sum_{i=1}^{u} P(l,c,u). \tag{9}$$

From this, the cumulative probability of observing from 1 to u unique alleles given c contributors, across all loci $l \in L$ is

$$C(c,u) = \prod_{l \in L} C(l,c,u). \tag{10}$$

Note that there is no need for the invocation of the inclusion-exclusion rule, since the probabilities are cumulative (see example in the next section). Then, the probability of observing u unique alleles at least once, across all loci, $P(c,\ u)$ (for $u > 1$) is given by:

$$P(c,u) = C(c,u) - C(c,u-1). \tag{11}$$

To aid in understanding this, consider the following example. Assume we know the cumulative probability $C(3,\ 4)$ of observing up to four unique alleles across all loci with three contributors, and the cumulative probability of $C(3,\ 3)$ of observing up to three unique alleles across all loci with three contributors. If we want to know the probability $P(3,\ 4)$, of observing four alleles, *at least once*, across all loci, with three contributors, we subtract the probability of observing 1 to 3 alleles from the probability of observing from 1 to 4 alleles. Thus we are left with a case where we have observed four alleles at least once, across all loci. Therefore, $P(3,\ 4) = C(3,\ 4) - C(3,\ 3)$, as in (11).

There are two conditions which must be met for this Probabilistic Mixture Model to be applicable:

- The features (loci) must be independent, or able to be treated as such for the purpose of evaluating all possible combinations.
- The combining function (i.e., *max*, *min*) must allow for the use of cumulative probabilities.

An exhaustive list of the possible combining functions which will work with this method is beyond the scope of this work.

2.2 Computational Complexity

For a given locus l, the number of unique alleles u can range from 1 to $2c$, where c is the number of contributors to the mixture (although, of course, u can never exceed $|l|$). Therefore, there are at most $\binom{|l|}{u}$ ways to pick these u unique alleles. Once the unique alleles have been chosen, choose d duplicate alleles, where $d = 2c - u$. There are,

$$\frac{(u+d-1)!}{(u-1)!d!}, \tag{12}$$

ways to choose d duplicates from u unique alleles (the standard formula for the number of combinations of u items taken d at a time, with replacement). Therefore, the computational complexity, over all loci $l \in L$, is:

$$O\left(\sum_{l \in L}\sum_{u=1}^{max(2c,|l|)} \binom{|l|}{u}\frac{(u+d-1)!}{(u-1)!d!}\right). \tag{13}$$

Thus the computational complexity depends on the cardinality of the alleles at each locus, and only indirectly on the number of profiles available for analysis. While adding more profiles to the data set may introduce new alleles, there is a quasi-stable upper bound on the number of commonly observed alleles at each locus, imposed by the

TABLE 1
Details Regarding the Various Data Sets Used for Comparison of the Two Algorithms

Dataset	Maximum Number of Unique Alleles	Number of Profiles	Number of Subpopulations
FBI	27	959	6
CFS	31	834	4
RCMP	16	500	4

The "Maximum Number of Unique Alleles" is the highest number of unique alleles seen at any locus in the data set. The exact number of subpopulations contained within the data sets from Australia and Ireland are not known.

timescale of biological mutations compared to the probable lifetime of this technology. Suppose that, over time, human mutations accumulate to the point where twice as many unique alleles are observed, per locus, than are currently displayed by the FBI data set. The time required to run the probabilistic model analysis only increases from 42 seconds to 4.3 hours (approximately). This also assumes that computers do not increase in speed over this same evolutionary time period, which seems unlikely [24].

2.3 Results

This Probabilistic Mixture Model was run on several different data sets, for single individuals through five-person mixtures. The data sets used are from the Federal Bureau of Investigation [25], the Toronto Centre of Forensic Sciences and Royal Canadian Mounted Police [26]. A short summary of the data sets is shown in Table 1. When an allele occurred rarely (less than five times out of all profiles), the frequency was set to $5/2n$, n being the number of profiles present in the data set [27]. The results are contrasted with the direct combinations method for 3-person mixtures described in [28], and a comparison of the resulting misclassification rates for 3-person mixtures is shown in Table 2

TABLE 2
Comparison of Misclassification Rates of Three-Person Mixtures between Experimental and Probabilistic Methods

Dataset	Experimental (*exp*)	Probabilistic (*pmm*)	Difference
FBI	3.39%	2.82%	20.38%
African American	4.00%	2.29%	74.44%
Bahamian	3.15%	1.75%	79.69%
Caucasian	4.80%	3.83%	25.23%
Jamaican	4.22%	2.32%	82.29%
S.W. Hispanic	5.97%	4.01%	48.91%
Trinidadian	2.88%	1.26%	128.93%
CFS	4.10%	2.96%	38.56%
Asian	5.43%	3.89%	39.77%
Black	4.17%	1.63%	156.62%
Caucasian	6.07%	5.05%	20.13%
East Indian	5.04%	3.12%	61.43%
RCMP	8.62%	6.40%	34.60%
Caucasian	5.14%	3.66%	40.51%
North Ontario Ab.	20.17%	14.81%	36.19%
Salishan Ab.	10.76%	8.96%	20.10%
Saskatchewan Ab.	14.84%	7.67%	93.43%

The difference is equal to (exp-pmm)/exp.

A sample of the output, obtained by running the algorithm on the FBI data set, is shown in Tables 3 and 4 for the probabilities and cumulative probabilities, respectively. Normally, one would expect the cumulative probabilities to sum to 1.0. However, since the allele frequencies were modified for rare alleles, this is not the case (the frequency of rare alleles was set to $5/2n$, and thus the allele frequencies do not sum to 1.0). Thus, when the computations are completed, the results must be normalized individually for each column (refer to Table 4). Although

TABLE 3
Probabilistic Mixtures on FBI Data Set, from 1 to 5 Contributors, Probability of Observing 1 to 2c Peaks

Maximum # of Observed Peaks	Number of Contributors (*c*)				
	1	2	3	4	5
1	$5.757E-10$	$4.779E-26$	$4.074E-41$	$8.258E-56$	$2.502E-70$
2	$1.000E+00$	$3.419E-09$	$4.038E-18$	$8.698E-27$	$3.616E-35$
3		$2.299E-02$	$5.424E-07$	$4.445E-12$	$3.469E-17$
4		$9.770E-01$	$2.816E-02$	$9.174E-05$	$1.372E-07$
5			$6.093E-01$	$8.415E-02$	$4.540E-03$
6			$3.625E-01$	$5.413E-01$	$2.027E-01$
7				$3.324E-01$	$4.958E-01$
8				$4.203E-02$	$2.522E-01$
9					$4.253E-02$
10					$2.266E-03$

TABLE 4
Probabilistic Mixtures on FBI Data Set, from 1 to 5 Contributors, Cumulative Probability of Observing 1 to 2c Peaks

Maximum # of Observed Peaks	Number of Contributors (*c*)				
	1	2	3	4	5
1	$5.757E-10$	$4.779E-26$	$4.074E-41$	$8.258E-56$	$2.502E-70$
2	$1.000E+00$	$3.419E-09$	$4.038E-18$	$8.698E-27$	$3.616E-35$
3		$2.299E-02$	$5.424E-07$	$4.445E-12$	$3.469E-17$
4		$1.000E+00$	$2.816E-02$	$9.174E-05$	$1.372E-07$
5			$6.375E-01$	$8.424E-02$	$4.540E-03$
6			$1.000E+00$	$6.255E-01$	$2.073E-01$
7				$9.580E-01$	$7.030E-01$
8				$1.000E+00$	$9.552E-01$
9					$9.977E-01$
10					$1.000E+00$

TABLE 5
Probabilistic Mixtures on FBI Data Set, from 1 to 5 Contributors, Cumulative Probability of Observing 1 to 2c Peaks, after Ignoring the Most Informative Locus

Maximum # of	Number of Contributors (c)				
Observed Peaks	1	2	3	4	5
1	$5.140E-09$	$2.428E-23$	$9.905E-37$	$8.949E-50$	$1.173E-62$
2	$1.000E+00$	$4.081E-08$	$6.920E-16$	$2.054E-23$	$1.107E-30$
3		$4.245E-02$	$5.570E-06$	$2.936E-10$	$1.490E-14$
4		$1.000E+00$	$6.097E-02$	$6.582E-04$	$3.665E-06$
5			$7.039E-01$	$1.699E-01$	$2.083E-02$
6			$1.000E+00$	$7.057E-01$	$3.500E-01$
7				$9.580E-01$	$7.562E-01$
8				$1.000E+00$	$9.552E-01$
9					$9.977E-01$
10					$1.000E+00$

the 3-way mixtures are further considered below, take note of the 4-contributor column. Over 75 percent of mixtures from four individuals will appear to be, by strict allele counting, to be comprised of three or less contributors.

2.3.1 Eliminating One Locus

A DNA analyst is often allowed to discard one locus that appears to be anomalous with respect to the number of contributors [29]. For instance, if a profile exhibits four or less alleles at 12 loci, but one locus has five alleles present, the analyst may disregard that 5-allele locus, and determine that the sample originated with only two individuals. It is not possible for this approach to disregard the locus at which the highest number of unique alleles is observed, as is done with the experimental three-way mixtures in [28]. This is implicit in the nature of the probabilistic approach, in that scenarios are constructed for a given number of unique alleles at each locus, and the results multiplied over all loci. However, it is possible to do something similar. When the cumulative probabilities are multiplied across all loci, it would be possible to skip the multiplication for the one locus providing the most change; this would be the locus whose cumulative probability is the lowest.

The results of performing this analysis on the FBI data set are shown in Table 5. The results from this table cannot be directly compared to those obtained in [28], since ignoring the locus with the highest number of visible alleles in every mixture is not the same as ignoring the most informative locus in the Probabilistic Mixture Model.

As there are 30 entries in the table, while analyzing the data, there are 30 occasions when it is necessary to disregard a locus. Of these 30, one locus is disregarded 25 times (or 83.33 percent of the time), D18S51. Of the other five occurrences, FGA is disregarded four times, and THO1 is disregarded once.

3 Determining the Number of Contributors

A fundamental goal of this research is to predict, given a sample DNA mixture from a crime scene, the most likely number of contributors to that mixture. The probabilistic algorithm developed above provides a novel approach.

3.1 Approach

Consider a sample which, by allele counting alone, appears to be a mixture of m or more individuals. Define the "unique" alleles for the probabilistic algorithm to be those observed in the sample. Select an appropriate data set of allele frequencies. From this point, the probabilistic algorithm can be used to pick, for each locus, all possible sets of duplicate alleles which will result in a complete mixture of m or more individuals. These probabilities, such as those shown in Table 3, can be used to produce likelihood ratios, which can then be used to probabilistically analyze the most likely actual number of contributors to the mixture.

For each allele observed in the sample, that allele is added to the number of alleles observed in the underlying data set, and the allele frequencies are adjusted, as described in [30]. However, for each allele in the sample that did not exist in the frequency data set, the user should be warned that they might wish to perform the analysis again, utilizing a different data set (in which that particular allele has been previously observed).

If the maximum number of peaks observed in the sample is p, then the minimum possible number of contributors is given by:

$$m = \left\lfloor \frac{p}{2} \right\rfloor. \tag{14}$$

For each $i,\ j$, where $j > i \geq m$, the likelihood ratio of this sample originating from i contributors versus j contributors is given by:

$$LR = \frac{P(i, 2m)}{P(j, 2m)}, \tag{15}$$

where the formula for $P(c,\ u)$ is that given in (11). In our example, comparing $i = 2$ contributors to $j = 3$, with a minimum number of contributors being $m = 2$, the values are thus taken from row 4 and columns 2 and 3 of Table 3. If the likelihood ratio is greater than one, then the sample is considered more likely to be a mixture of two individuals. If LR is less than one, then the sample is considered more likely to be a mixture of three individuals.

It is also possible, for testing purposes, to produce all possible two-way or three-way mixtures. Each mixture thus constructed can be used as a sample input as described above. However, unlike the single sample file approach, in this mode, the "sample" alleles should not be readded to the data set (and thus would not change the allele frequencies). The mixtures were generated by using the profiles present in the listed data sets. For instance, the FBI Caucasian data set [22] contains the actual, anonymous profiles of the 194 individuals who contributed to the U.S. Caucasian allele frequencies. There are thus 18,721 2-way mixtures, all of which were used. There are also 1,198,144 3-way mixtures, of which 57,569 (the 4.80 percent seen in Table 2) appeared to be 2-way mixtures via allele counting. These 57,569 mixtures are the ones used for analysis.

3.1.1 Compensating for Substructure

It has been observed that the frequency of homozygotes in human populations is often slightly greater than what would have resulted from chance associations of alleles. The value most common practice used in forensic casework to compensate for this observation is to multiply by a conservatively selected constant, referred to as θ. The NRC II report suggests a value of 0.01 as being conservative [27]. The NRC II report also suggests two formulae for use with homozygous and heterozygous loci (pg. 102), respectively:

$$A_iA_i : P_{ii} = p_i^2 + p_i(1-p_i)\theta_{ii}, \tag{16}$$

$$A_iA_j : P_{ij} = 2p_ip_j(1-\theta_{ij}). \tag{17}$$

Using the method described above would produce an intractable computational problem: instead of considering the combination of alleles present in a mixture, it would be necessary to examine every permutation for the presence of homozygotes. Further, inbreeding and inflated homozygosity are not the only issue: the allele frequencies used are only estimates. One must consider the possibility that the population of interest has diverged from the reference population. The more that individuals are related to each other, the more they (and their allele frequencies) have diverged from the population as a whole. However, a more general approach which considers structured populations [31] both considers the increased importance of θ and makes the analysis of all permutations of the alleles unnecessary. The following example, taken from [31], shows the probability of observing a sample containing the alleles a, b, c, and c:

$$Pr(abcc) = \frac{[(1-\theta)p_a][(1-\theta)p_b][(1-\theta)p_c][(1-\theta)p_c+\theta]}{(1-\theta)(1)(1+\theta)(1+2\theta)}. \tag{18}$$

Thus one can see that, in (18), the denominator only depends upon the total number of observed alleles, while the numerator depends upon how many times each allele is observed. Suppose U is the set of unique alleles observed and that the number of unique alleles is $|U|$. If n_i is the number of times that allele u_i is observed, then the total number of alleles (including duplicates) $n = \sum_{i=1}^{|U|} n_i$. If p_i is the frequency of that same allele, a more general form of (18) is

$$P = \frac{\prod_{i=1}^{|U|}\prod_{j=1}^{u_i}[(1-\theta)p_{u_i}+\theta(j-1)]}{\prod_{i=1}^{n} 1+\theta(i-2)}, \tag{19}$$

where $n=4$, $U=\{a,b,c\}$, $|U|=3$, $u_1=a$, $u_2=b$, $u_3=c$, $n_1=n_2=1$, and $n_3=2$. Equation 3 can be restated by multiplying (19) by $perms$:

$$P = \frac{\prod_{i=1}^{|U|}\prod_{j=1}^{u_i}[(1-\theta)p_{u_i}+\theta(j-1)]}{\prod_{i=1}^{n} 1+\theta(i-2)} \cdot perms. \tag{20}$$

TABLE 6
Percent of Actual Two-Person Mixtures, Correctly Identified as Being a Two-Way Mixture by the Probabilistic Mixture Model

Dataset	Correctly Identified
FBI	99.07%
African American	98.08%
Bahamian	99.04%
Caucasian	98.48%
Jamaican	98.01%
S.W. Hispanic	98.10%
Trinidadian	98.88%
CFS	99.14%
Asian	97.87%
Black	98.91%
Caucasian	98.24%
East Indian	98.72%
RCMP	98.54%
Caucasian	97.97%
North Ontario Ab.	95.37%
Salishan Ab.	96.58%
Saskatchewan Ab.	97.72%

The remainder of the analyses presented herein were calculated using (20) with $\theta = 0.01$, and enforcing $5/2n$ as the minimum frequency of an allele (where n is the number of profiles present in the data set).

3.2 Results

For each two-way mixture generated, we calculate the likelihood ratio. If the LR is > 1.0, the analysis predicts that two contributors is more likely than three (a correct classification). Otherwise ($LR \leq 1.0$), the mixture has been identified as originating with a two-person mixture (an incorrect classification).

For each three-way mixture generated that could potentially be mistaken as a two-person mixture (the maximum number of alleles observed across all loci is four or less), we calculate the likelihood ratio. If the LR is greater than 1.0, the analysis predicts that two contributors is more likely than three (an incorrect classification). Otherwise ($LR \leq 1.0$), the mixture has been identified as originating with a three-person mixture (a correct identification).

Tables 6 and 7 present the results of applying this approach to several data sets. Utilizing only the allele frequencies, over 95 percent of two-contributor samples are correctly classified as having originated with a mixture of the DNA of two individuals (Table 6). In most of the analyzed data sets, the classification rate is better than 98 percent. This shows that the use of this approach will seldom send investigators searching for a contributor who does not in fact exist.

Of those three-contributor samples which would initially appear to originate with two contributors, over 68 percent are correctly classified as having come from a mixture of three contributors (Table 7). Using the cognate allele frequencies results in improved classification rates, relative to allele frequencies from the combined data set. Without the use of this approach, there may be scant evidence to suggest to investigators that additional contributor(s) to the sample may exist.

If the likelihood ratios generated in the analysis of 2- and 3-way mixtures of the FBI data set are binned according to

TABLE 7
Percent of Actual Three-Person Mixtures, with Only four or Less Alleles Apparent at Every Locus, Correctly Identified as Being a Three-Way Mixture by the Probabilistic Mixture Model

Dataset	Correctly Identified
FBI	70.49%
African American	81.01%
Bahamian	73.11%
Caucasian	78.74%
Jamaican	80.92%
S.W. Hispanic	78.43%
Trinidadian	78.92%
CFS	69.87%
Asian	79.56%
Black	72.09%
Caucasian	78.34%
East Indian	76.61%
RCMP	68.92%
Caucasian	79.60%
North Ontario Ab.	82.10%
Salishan Ab.	81.48%
Saskatchewan Ab.	80.30%

the $log_{10}(LR)$, the result can be shown as a histogram. For the 2-way mixtures, a $LR > 1$ indicates that the mixture has been correctly classified as originating with two contributors. This corresponds to $log_{10}(LR) \geq 0$, as in Fig. 2. For the 3-way mixtures, a $LR < 1$ indicates correct classification as originating with three contributors, and thus corresponds to $log_{10}(LR) < 0$ in Fig. 3.

The three-person mixtures generated during this process are the same that are generated in [28]. That approach yields 3.39 percent of the three-person mixtures that are mischaracterized as originating with two individuals. However, the application of the Probabilistic Mixture Model eliminates over 70 percent of those misclassifications. Thus the misclassification rate is effectively lowered from 3.39 to 1.00 percent.

By varying the LR threshold at which the decision is made, the class discrepancy seen in Section 3.2 can be overcome. The result of using various LR thresholds is shown in Fig. 4. The point where the curves cross (and class balance is achieved) is near 93 percent, and is achieved by utilizing a LR threshold of approximately 13.

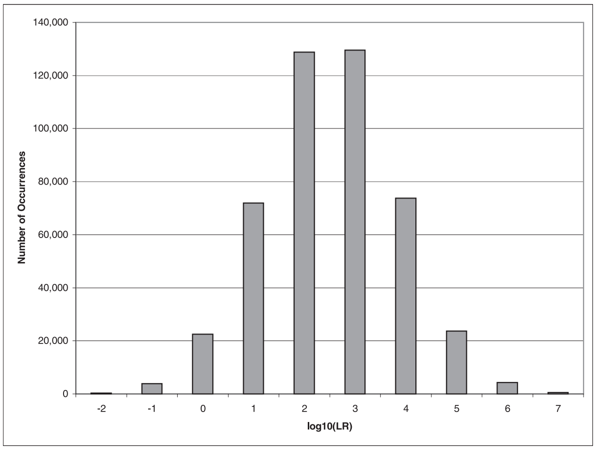


Fig. 2. Histogram for the likelihood ratios of 2-way mixtures from the FBI data set.

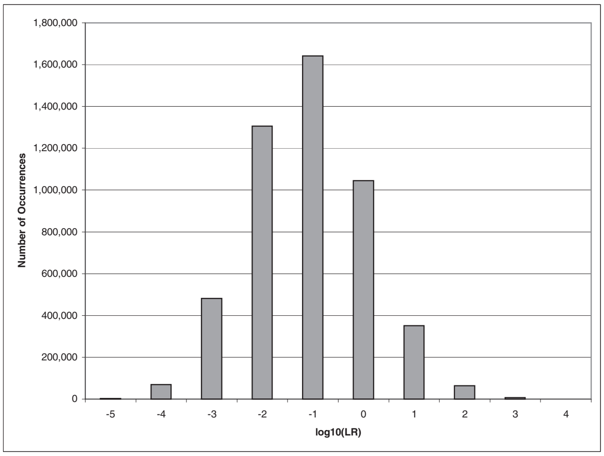


Fig. 3. Histogram for the likelihood ratios of 3-way mixtures from the FBI data set.

3.3 Characterizing Samples

It proves informative to consider the makeup of samples which will generate likelihood ratios greater than 1, and compare them to those which generate LR less than one. Table 8 shows a profile with exactly four alleles observed at each locus, these four alleles being the most commonly occurring allele at that locus (within the FBI data set). While the alleles themselves are of little interest, the makeup of this profile is. It yields $LR = 2.933E - 6$, which means that, if the number of contributors is actually 3, it is $3.410E + 05$ times more likely to observe this profile than it is if there the true number of contributors is two. Simply observing that the maximum number of alleles across all loci is four might lead one to suppose that the sample originates from two or more individuals.

Now let us consider a similar profile, in which we simply eliminate the fourth allele (the fourth most common) from each locus. For this sample, $LR = 1.088E + 01$, meaning that this sample is approximately 10 times more likely to be observed if it had originated from two contributors than if it originated with three contributors. This makes sense, if one

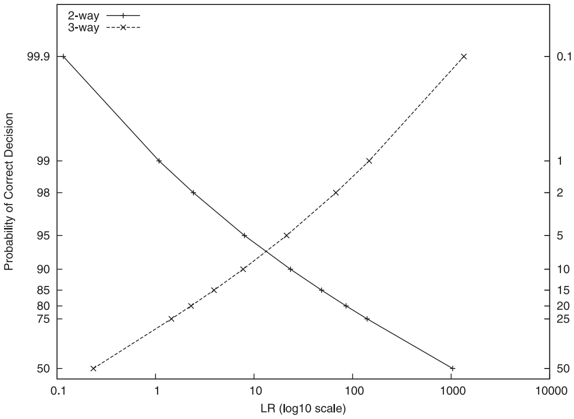


Fig. 4. Effect of using various LR thresholds on accuracy. The 2-way mixtures (solid line) are known to be 2-way mixtures, thus a correct classification is when the $LR > 1$. The 3-way mixtures (dashed line) are known to be 3-way mixtures but appear, via allele counting, to be mixtures of two individuals; thus a correct classification is when the $LR < 1$. The probabilities on the y-axis are plotted on a logit scale: $p\%$ is plotted at the point $log_{10}(p/(100 - p))$.

TABLE 8
Sample STR Mixture Exhibiting the four Most Commonly Occurring Alleles at Each Locus (According to The FBI Data Set)

Locus	Sample Mixture			
D3S1358	15,	16,	17,	14
vWA	16,	17,	18,	15
FGA	22,	24,	23,	21
D8S1179	14,	13,	15,	12
D21S11	30,	28,	29,	32.2
D18S51	17,	16,	15,	14
D5S818	12,	11,	13,	10
D13S317	12,	11,	13,	9
D7S820	10,	11,	12,	9
CSF1PO	12,	10,	11,	13
TPOX	8,	11,	9,	10
THO1	7,	6,	8,	9.3
D16S539	11,	12,	9,	13

stops to consider what would have to occur for two individuals to create the 4-alleles profile. Both individuals would have to be heterozygous at all 13 loci, and would have to have no alleles in common at any locus. On the other hand, for two individuals to create the second sample (three alleles at each locus), they could both be heterozygous, and could each share one of their two alleles with the other individual, or one could be heterozygous and the other homozygous, and they share no alleles. Either situation could occur at each of the 13 loci, as long as the result is the observation of three unique alleles. There are many more ways this could happen, and thus the overall probability of this happening is higher than the probability of both individuals being heterozygous at every locus, and having no alleles in common at every locus.

What happens if one selects the 3 or 4 most rare alleles at every locus, instead of the most common? In these cases, two contributors are much more likely than three contributors: with exactly three alleles observed at every locus, it is $1.322E+28$ times more likely to be observed with two contributors; with four alleles at every locus, it is $1.162E+20$ times more likely. This also make sense: the probability of two individuals who are heterozygous at every locus has become greater than the probability of observing three individuals, all of whom have the most rare alleles at every locus.

When examining the experimental 2-way mixtures from the FBI data set which, by likelihood ratio, are classified as 3-way mixtures, it is observed that they are similar to the above. Of the seven lowest LR (all those below 0.01), most loci share four alleles, while several share three. The average number of unique alleles per locus, across these seven profiles, is 9.14.

3.4 Bayesian Concerns

Ideally, one would like to be able to state how many individuals have contributed to a DNA sample, along with a degree of confidence in that conclusion. The methodology presented here does not, by itself, fully address this overarching question. This approach measures how well the observed data fits any specific hypothesis for the number of contributing individuals. In order to answer the overarching question, one must also know how likely each competing hypothesis is a priori.

Assume, for the purposes of illustration, that we examine a sample and determine that it is either a relatively rare mixture of 2-contributors or an equally rare mixture of 3-contributors. Given that the sample is equally rare for either conclusion, we have essentially a 50 percent chance of making an error. Let us further assume, that, in practice, mixtures of two contributors are encountered 20 times more often than mixtures of three persons. Under this assumption, we could conclude that our sample is a mixture of two contributors with a greatly reduced likelihood of error. Essentially, skewing the odds of any particular mixture being more likely to be a mixture of two contributors prior to examining the data just due to relative frequency (prior odds) proportionally reduces the effective relative rarity for every mixture of two contributors in a likelihood calculation.

This problem is by no means unique to forensics. Such Bayesian concerns [32] plague many statistical approaches. Such concerns are easily dealt with when the prior odds for each competing hypothesis are known. Herein, we have assumed that competing hypothesizes are a priori equally likely. In actual forensic work, the prior odds must be specifically (and empirically) addressed.

4 Conclusion

The Probabilistic Mixture Model, described above, is almost 99 percent accurate at correctly identifying mixtures where the number of contributors could also have been inferred by counting the maximum number of alleles observed at any of the loci for which information is available. When the presence of one contributor is masked the method correctly identifies the actual number of contributors over 68 percent of the time (a significant improvement over allele counting approaches which would be invariably incorrect). An Applet implementing the algorithm is available online for public use at http://www.personal.psu.edu/drp15/tools/pmm/. This method could also be combined with other approaches (such as those that rely upon peak height or area) to form a consensus. The Probabilistic Mixture Model could also be further adapted to samples where the DNA profile of one of the contributors is known, such as a victim. This would allow for more accurate determination of the most likely total number of contributors to the sample.

Appendix

Hypothetical Crime Scene Sample

Consider the hypothetical crime scene sample shown in Table 9. Observing three to four alleles at each locus, one might believe that the sample originated with two individuals. However, this mixture was created from three profiles found in the FBI's U.S. Caucasian data set, labeled C003, C119, and C160. The sum of the probabilities of all 2-contributor explanations is $3.014E-23$, while the probability of three contributors is $1.287E-18$. While neither probability is very large, the likelihood ratio indicates that a 3-contributor explanation for this sample occurs 42,709 times more often than does a two-contributor explanation. These values are achieved using the U.S. Caucasian data set. If one did not know the racial origin of the samples, one could

TABLE 9
Hypothetical Crime Scene Sample

Locus	Sample Mixture			
D3S1358	14,	15,	17,	18
vWA	15,	16,	17,	18
FGA	18,	21,	23,	24
D8S1179	10,	12,	13,	14
D21S11	29,	30,	31,	31.2
D18S51	13,	14,	16,	17
D5S818	10,	11,	12,	13
D13S317	10,	11,	12,	14
D7S820	8,	10,	11,	12
CSF1PO	10,	11,	12,	13
TPOX	8,	9,	11	
THO1	6,	7,	9,	9.3
D16S539	9,	11,	12,	14

use the combined FBI data set, which indicates that a three-contributor explanation is 11,308 times more likely to occur.

References

[1] J. Travis and R. Rau, “Convicted by Juries, Exonerated by Science: Case Studies in the Use of DNA Evidence to Establish Innocence after Trial,” Technical Report NCJ 161258, US Dept. of Justice, http://www.ncjrs.org/pdffiles/dnaevid.pdf, June 1996.

[2] J.D. Watson and F.H. Crick, “Genetical Implications of the Structure of Deoxyribonucleic Acid,” *Nature,* vol. 171, pp. 964-967, 1953.

[3] J.D. Watson and F.H. Crick, “Molecular Structure of Nucleic Acids: A Structure for Deoxynucleic Acids,” *Nature,* vol. 171, pp. 737-738, 1953.

[4] A. Edwards, H.A. Hammond, J. Lin, C.T. Caskey, and R. Chakraborty, “Genetic Variation at Five Trimeric and Tetrameric Tandom Repeat Loci in Four Human Population Groups,” *Genomics,* vol. 12, pp. 241-253, 1992.

[5] C.J. Frégeau and R.M. Fourney, “DNA Typing with Flourescently Tagged Short Tandom Repeats: A Sensitive and Accurate Approach to Human Identification,” *BioTechniques,* vol. 15, pp. 100-119, 1993.

[6] T.R. Moretti, A.L. Baumstark, D.A. Defenbaugh, K.M. Keys, A.L. Brown, and B. Budowle, “Validation of STR Typing by Capillary Electrophoresis,” *J. Forensic Sciences,* vol. 46, no. 3, pp. 661-676, 2001.

[7] C.J. Frégeau, K.L. Bowen, and R.M. Fourney, “Validations of Highly Polymorphic Fluorescent Multiplex Short Tandem Repeat Systems Using Two Generations of DNA Sequencers,” *J. Forensic Sciences,* vol. 44, no. 1, pp. 133-166, 1999.

[8] A.T. Woolley and R.A. Mathies, “Ultra-High-Speed DNA Fragment Separations Using Microfabricated Capillary Array Electrophoresis Chips,” *Proc. Nat'l Academy of Sciences USA,* vol. 91, no. 24, pp. 11348-11352, 1994.

[9] J.R. Gilder, T.E. Doom, K. Inman, and D.E. Krane, “Run-Specific Limits of Detection and Quantitation for STR-Based DNA Testing,” *J. Forensic Sciences,* vol. 52, no. 1, pp. 97-101, 2007.

[10] J.M. Butler, *Forensic DNA Typing,* second ed. Elsevier Academic Press, 2005.

[11] J.R. Gilder, T.E. Doom, M.L. Raymer, K.G. Inman, and D.E. Krane, “Resolution of Forensic DNA Mixtures,” submitted to *Forensic Science Int'l: Genetics,* Oct. 2008.

[12] S.L. Lauritzen and J. Mortera, “Bounding the Number of Contributors to Mixed DNA Stains,” Technical Report R-02-2003, Dept. of Mathematical Sciences, Aalborg University, http://www.mathnet.or.kr/papers/Aalborg/Steff/netshort.ps, Feb. 2002.

[13] M.W. Perlin and B. Szabady, “Linear Mixture Analysis: A Mathematical Approach to Resolving Mixed DNA Samples,” *J. Forensic Sciences,* vol. 46, no. 6, pp. 1372-1377, 2001.

[14] T. Wang, N. Xue, M. Rader, and J.D. Birdwell, “Least Square Deconvolution (LSD) of STR/DNA Mixtures,” *Proc. Seventh CODIS User's Conf.,* Oct. 2001.

[15] P. Gill, R. Sparkes, R. Pinchin, T. Clayton, J. Whitaker, and J. Buckleton, “Interpretating Simple STR Mixtures Using Allele Peak Areas,” *Forensic Science Int'l,* vol. 91, pp. 41-53, 1998.

[16] R.G. Cowell, S.L. Lauritzen, and J. Mortera, “Identification and Separation of DNA Mixtures Using Peak Area Information,” *Forensic Science Int'l,* vol. 166, no. 1, pp. 28-34, 2007.

[17] R.G. Cowell, S.L. Lauritzen, and J. Mortera, “A Gamma Model for DNA Mixture Analyses,” *Bayesian Analysis,* vol. 2, no. 2, pp. 333-348, 2007.

[18] J.R. Gilder, K.G. Inman, W.M. Shields, and D.E. Krane, “Magnitude-Dependant Variation in Peak Height Balance at Heterozygous STR Loci,” *Int'l J. Legal Medicine,* vol. 125, no. 1, pp. 87-94, 2011.

[19] J. Mortera, A.P. Dawid, and S.L. Lauritzen, “Probabilistic Expert Systems for DNA Mixture Profiling,” *Theoretical Population Biology,* vol. 63, pp. 191-205, 2003.

[20] R.G. Cowell, S.L. Lauritzen, and J. Mortera, “Probabilistic Modelling for DNA Mixture Analysis,” *Forensic Science Int'l: Genetics Supplement Series; Progress in Forensic Genetics 12—Proc. 22nd Int'l ISFG Congress,* vol. 1, no. 1, pp. 640-642, Aug. 2008.

[21] T. Egelund, I. Dalen, and P.F. Mostad, “Estimating the Number of Contributors to a DNA Profile,” *Int'l J. Legal Medicine,* vol. 117, no. 5, pp. 271-275, 2003.

[22] B. Budowle, T.R. Moretti, A.L. Baumstark, D.A. Defenbaugh, and K.M. Keys, “Population Data on the Thirteen CODIS Core Short Tandem Repeat Loci in African Americans, U.S. Caucasians, Hispanics, Bahamians, Jamaicans, and Trinidadians,” *J. Forensic Sciences,* vol. 44, no. 6, pp. 1277-1286, 1999.

[23] R. Chakraborty and K.K. Kidd, “The Utility of DNA Typing in Forensic Work,” *Science,* vol. 254, pp. 1735-1739, 1991.

[24] G.E. Moore, “Cramming More Components onto Integrated Circuits,” *Electronics,* vol. 38, no. 8, pp. 114-117, 1965.

[25] B. Budowle and T.R. Moretti, “Genotype Profiles for Six Population Groups at the 13 CODIS Short Tandem Repeat Core Loci and Other PCR-Based Loci,” vol. 1, no. 2, http://www.fbi.gov/hq/lab/fsc/backissu/july1999/dnaloci.txt, July 1999.

[26] Population Studies Data Centre, “Raw Data,” *Canadian Soc. Forensic Science,* http://www.csfs.ca/pplus/profiler.htm, 2006.

[27] National Resource Council, *The Evaluation of Forensic DNA Evidence.* National Academy Press, 1996.

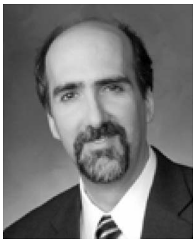
[28] D.R. Paoletti, T.E. Doom, C.M. Krane, M.L. Raymer, and D.E. Krane, “Empirical Analysis of the STR Profiles Resulting from Conceptual Mixtures,” *J. Forensic Sciences,* vol. 50, no. 6, pp. 1361-1366, 2005.

[29] The New York Office of the Chief Medical Examiner, *Forensic Biology Protocols for Forensic STR Analysis,* June 2008.

[30] D.J. Balding and R.A. Nichols, “DNA Profile Match Probability Calculation: How to Allow for Population Stratification, Relatedness, Database Selection and Single Bands,” *Forensic Science Int'l,* vol. 64, pp. 125-140, 1994.

[31] J.M. Curran, C.M. Triggs, J. Buckleton, and B.S. Weir, “Interpreting DNA Mixtures in Structured Populations,” *J. Forensic Sciences,* vol. 44, no. 5, pp. 987-995, 1999.

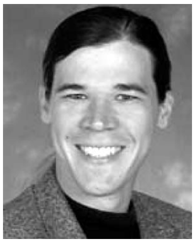
[32] T. Bayes, “Studies in the History of Probability and Stistics: IX. Thomas Bayes' Essay towards Solving a Problem in the Doctrine of Chances,” *Biometrika,* vol. 45, pp. 296-315, 1763.



David Paoletti received the BS degrees in computer science and electrical engineering from Michigan Technological University, Houghton, in 1990, the MS degree in computer science from Michigan State University, East Lansing, in 1992, and the PhD degree in computer science from Wright State University, Dayton OH, in 2006. He is assistant professor at Pennsylvania State University Beaver. His research interests include bioinformatics, evolutionary computation, and population genetics. He is a member of the IEEE.

Dan Krane received the PhD degree. He is currently working as a professor in the Department of Biological Sciences at Wright State University. He helped develop Genophiler and is founder and president of Forensic Bioinformatic Services, Inc. A leading authority on forensic DNA evidence, he has testified as an expert witness in more than 75 cases.

Michael Raymer received the BS degree in computer science from Colorado State University, Fort Collins, in 1991, the MS degree in computer science from Michigan State University, East Lansing, in 1995, and the PhD degree in computer science from Michigan State University in 2000. He is associate professor at Wright State University, where he is also a member of the Biomedical Sciences Program faculty. His research interests include evolutionary computation, pattern recognition, computational biology, protein structure and function, and bioinformatics. He is cofounder of the company Forensic Bioinformatic Systems, which performs computational analysis of forensic DNA evidence. He is a member of the IEEE.



Travis Doom received the MSc and PhD degrees from Michigan State University, East Lansing. He is associate professor at Wright State University (WSU), Dayton, OH in the Department of Computer Science and Engineering. He also holds positions in the Department of Electrical Engineering and is a member of the graduate faculty in Biomedical Sciences. He is codirector of the bioinformatics research group at WSU and a cofounder of Forensic Bioinformatics Services, a company which performs computational analysis of forensic DNA evidence. His primary areas of teaching and research interest include digital systems, bioinformatics, computer architecture and operating systems, design automation, and computational mathematics/theory. He is a senior member of the IEEE.