BIOINFORMATICS

Vol. 22 no. 14 2006, pages e298–e306
doi:10.1093/bioinformatics/btl200

Interpreting anonymous DNA samples from mass disasters—probabilistic forensic inference using genetic markers

Tien-ho Lin[1], Eugene W. Myers[2] and Eric P. Xing[1,*]

[1]School of Computer Science, Carnegie Mellon University, Pittsburgh, PA and [2]HHMI Janelia Farms Research Campus, Ashburn, VA

ABSTRACT

Motivation: The problem of identifying victims in a mass disaster using DNA fingerprints involves a scale of computation that requires efficient and accurate algorithms. In a typical scenario there are hundreds of samples taken from remains that must be matched to the pedigrees of the alleged victim's surviving relatives. Moreover the samples are often degraded due to heat and exposure. To develop a competent method for this type of forensic inference problem, the complicated quality issues of DNA typing need to be handled appropriately, the matches between every sample and every family must be considered, and the confidence of matches need to be provided.

Results: We present a unified probabilistic framework that efficiently clusters samples, conservatively eliminates implausible sample-pedigree pairings, and handles both degraded samples (missing values) and experimental errors in producing and/or reading a genotype. We present a method that confidently exclude forensically unambiguous sample-family matches from the large hypothesis space of candidate matches, based on posterior probabilistic inference. Due to the high confidentiality of disaster DNA data, simulation experiments are commonly performed and used here for validation. Our framework is shown to be robust to these errors at levels typical in real applications. Furthermore, the flexibility in the probabilistic models makes it possible to extend this framework to include other biological factors such as interdependent markers, mitochondrial sequences, and blood type.

Availability: The software and data sets are available from the authors upon request.

Contact: epxing@cs.cmu.edu

1 INTRODUCTION

Rapid advances in genotyping technology and mathematical theories of pedigrees have enabled their application in traditional forensic applications such as victim or perpetrator identification and paternity testing common place, even when family structures are complex or sample mixtures and mutations are involved (Mortera *et al*., 2003). A natural next step is to enlarge the scale of genetic forensic inference to mass disasters, such as airplane crashes, terrorist bombings, or battlefields, in which hundreds or even thousands of remains, usually highly degraded, have to be identified for all the victims according to DNA evidences from candidate family members (Egeland *et al*., 2000; Lauritzen and Sheehan, 2003). In addition to issues related to the increased scale of the problem, such a problem also poses new technical challenges such as the presence of errors in the genotypes and pedigrees, incomplete genetic information, and the need for decision making with very high confidence. (This last issue is typical of forensic cases, where seemingly low probability event such as incorrect victim/family matching can have serious legal consequence, and must be determined with a confidence much more stringent than usually adopted in experimental biology.)

DNA typing has long been used in forensic investigations, but until a decade ago, mass disaster victim identification has generally relied on dental and medical records, fingerprints, and even photographic evidence and personal effects (Ballantyne, 1997). These techniques require comparison between *ante mortem* (AM) information for the victim and *post mortem* (PM) information of the remains. However, in most mass disaster scenarios, AM information is not available for all victims and bodies are not intact, rendering such methods ineffective. Whitaker *et al*. (1995) established the use of short tandem repeat (STR) typing, or microsatellite markers, in mass disaster identification, and Olaisen *et al*. (1997) applied it to victim identification in the 1996 Spitsbergen aircraft accident, in which it proved to be highly reliable. A thirteen STR loci fingerprint set called the Combined DNA Index System (CODIS) is now in routine usage by the FBI, and has become a major tool in difficult disaster victim identification cases (Hsu *et al*., 1999; Cash *et al*., 2003).

While the basic problem of computing the likelihood ratio that a given sample is part of a given pedigree versus the null hypothesis of a random sample has been extensively studied (Olaisen *et al*., 1997), the inference problem of matching many pedigrees against many samples has not. Specialized software tools have been developed for large scale mass disaster identification (Cash *et al*., 2003) including the use of mitochondrial DNA (mtDNA) and single nucleotide polymorphism (SNP), but the matching algorithms utilized only rank the likely samples for each victim, and rank the likely victims for each sample. The complex interactions of all family evidence and all samples are not explored, and a great amount of expert involvement is still required. Moreover there is currently no systematic solution that addresses all the complicating factors: body part clustering, arbitrary pedigrees and their vetting, experimental genotyping error for the samples, partial genotypes due to heat and pressure damage of the DNA, and confidence of a cluster to family match based on other likely and

*To whom correspondence should be addressed.

unlikely family. This paper presents an architecture for the problem and a probabilistic framework that incorporates these uncertainties and scales to the required problem sizes.

We consider the following problem. We are given N family pedigrees for which the genotypes for some members are known, and the (potentially partial) genotypes of M samples belonging to the victims of a mass disaster. The problem is to match, with high confidence, the samples to the variable nodes (the purported victim reported by the family) of the pedigrees. Furthermore, we address how to screen out unambiguous matching outcomes and extract the truly ambiguous cases that merit costly personalized forensic investigation.

We approach the problem in two phases. First the samples are clustered into groups that have the same genotype. This reduces the problem of matching M samples to N pedigrees, to a smaller one of matching $J \ll M$ sample clusters to N pedigrees. During clustering possible errors in the STR data must be considered, especially when the DNA is degraded or when thousands of genotypes have been collected. We include a model for the types of errors that can occur in our probabilistic framework. In second phase, the cluster samples are matched to the variable nodes in the pedigrees. Forensic conclusions must be satisfactory from a legal perspective, as the purpose is to confirm the death of the victim, to return the remains to the families for closure, and in some cases to identify some of the victims as the perpetrators (in the case of terrorist acts). Therefore one can only make conclusions if there is a very small probability, typically 10^{-6} or smaller, of being wrong. We present a method to calculate the confidence of a certain match considering its likelihood ratio and other competitors for the slot. Then a forensically impossible match can be removed with high confidence.

Due to high confidentiality in disaster DNA data, simulation experiment is commonly performed so that true identity is known. We run three experiments with different simulation settings, and show that our algorithm is robust even with a lot of missing information and noise.

2 PRELIMINARIES

Consider M forensic samples from a mass disaster scene. Let $\mathbf{s}_1$, $\mathbf{s}_2$,...$\mathbf{s}_M$ denote the set of *sample genetic states* (to be specified shortly) retrieved from the M DNA samples, each from one of the forensic samples. Suppose there are N families that have filed missing person reports regarding this case (for presentation simplicity, we assume each family reports only one missing person, although generalization to multiple missing persons is feasible with our approach presented in the following), and have donated DNA samples as genetic references for victim identification. Let $\mathbf{f}_1$, $\mathbf{f}_2$ ... $\mathbf{f}_N$ denote the set of *familial genetic states* (defined in the sequel) obtained from these families.

Typically, body remains from a mass disaster and samples from donors are genetically characterized by a standard profile of K *microsatellite markers*. Each allele of such a marker corresponds to a numerical (in fact, discreet) reading from an electrophoresis gel; formally, we define each marker to be a random variable, and each of its alleles to be one of the realized states of this variable. For a forensic sample j, its sample genetic state (SGS) $\mathbf{s}_j \equiv (s_{j1}, s_{j2}, \ldots, s_{jK})$ denote the *genotype* profile of K markers, where $s_{jk} \equiv (s_{jk}^0, s_{jk}^1)$ represents an unordered pair of alleles of marker k from sample j. There are two alleles for each marker as human somatic cells are diploid, that is, there is a copy of a chromosome inherited from each parent. The superscripts ''1'' and ''0'' correspond to the parental origin of the alleles, i.e., paternal and maternal. Similarly, for each donor, we define $\mathbf{d}_j \equiv (d_{j1}, d_{j2}, \ldots, d_{jK})$ to be his/her genotype profile. Each family, say family i, may have multiple donors related by a *pedigree* T_i, therefore the familial genetic state (FGS) of a family with n_i donors is denoted by $\mathbf{f}_i \equiv \{\mathbf{d}_1, \ldots, \mathbf{d}_{n_i}; \mathbf{T}_i\}$. In typical mass disaster scenarios, multiple forensic samples (e.g., body remains) may belong to the same victim; therefore the samples can be grouped into clusters: i.e., $\mathbf{s}_1, \mathbf{s}_2, \ldots \mathbf{s}_M \Rightarrow \mathbf{c}_1, \mathbf{c}_2, \ldots \mathbf{c}_J$, where $\mathbf{c}_j = (\mathbf{c}_{j1}, \ldots, \mathbf{c}_{jm_j})$ and m_j denotes the size of cluster j (for simplicity, in the sequel we overload the symbol $\mathbf{c}_j$ to also represent the set of indices of SGSs belonging to cluster j). The forensic inference problem we concern here is that of determining the number of victims in the disaster, and the correct mapping between the victims and the reporting families.

In forensic applications, the microsatellite markers are chosen to be independent from each other (e.g., on different chromosomes). Via population censuring, the *a priori* probability (i.e., population frequency) of every allele of a microsatellite marker can be determined. Thus, given no familial information, the probability of an SGS of a forensic sample can be defined by the product of marker-specific genotype probabilities (by assuming the alleles are random samples from the population):

$$p(\mathbf{s}_j) = \prod_{k=1}^{K} p(s_{jk}), \tag{1}$$

where

$$p(s_{jk}) = \begin{cases} (\pi_{k,s_{jk}^0})^2 & \text{if} s_{jk}^0 = s_{jk}^1 \\ 2\pi_{k,s_{jk}^0}\pi_{k,s_{jk}^1} & \text{if} s_{jk}^0 \neq s_{jk}^1 \end{cases},$$

and $\pi_{k,a}$ denotes the population frequency of allele a of marker k.

The dependencies among donors from a family are captured by a pedigree. In our current setting, we consider only sexual inheritance among family members (i.e., donors plus the purported victim), and leave out nonsexual inheritance such as the mitochondria inheritance (incorporating such information is feasible in our framework and will be pursued in future research.) As illustrated in §3.4, a pedigree can be used to define the probability of the FGS of a family via a *probabilistic graphical model* (Pearl, 1988; Cowell *et al*., 1999). Note that a pedigree contains members who are not donors, nor victims, in order to specify the relations between the donors and the victim. These members represent the hidden variables in the graphical model, and will be marginalized out when computing the the FGS probability. For example, when the donor is the victim's brother, parents must appear on the pedigree even though their DNA samples are not available. The pedigree may have arbitrary structures, which are assumed to be correct after passing the validity check.

3 BODY IDENTIFICATION

To formulate a likelihood-ratio matching criteria for body identification, let's first assume that we have N reporting families and J victims (J will be determined by sample clustering as described in §3.2), and $J = N$. That is, each family has exactly one victim which corresponds to one cluster; and there is a one-to-one

alignment between the family pedigrees and the sample clusters. Our goal here is to find the optimal matching between $\{\mathbf{c}_j\}$ and $\{\mathbf{f}_i\}$. We will discuss how to relax the ''$J = N$'' and ''one-to-one correspondence'' assumptions later.

3.1 Matching via likelihood ratio

The matching between families and sample clusters can be represented by an $N \times N$ matching matrix $\mathbf{z}$, of which an element z_{ij} indicates the matching status between sample cluster j and family i:

$$z_{ij} = \begin{cases} 1 & \textit{if } \mathbf{c}_j \text{ is assigned to } \mathbf{f}_i \\ 0 & \text{otherwise} \end{cases}.$$

In case of one-to-one matching, $\mathbf{z}$ must satisfy the following constraints:

$$\sum_{i=1}^{N} z_{ij} = 1 \quad \forall j, \qquad \sum_{j=1}^{J} z_{ij} = 1 \quad \forall i. \tag{2}$$

Let $\pi(\mathbf{c}_j \mid \mathbf{f}_i)$ denotes the conditional probability of a cluster given a *matching* family, $\pi(\mathbf{c}_j)$ denotes the marginal probability of a cluster given no matching, and $p(\mathbf{f}_i)$ denotes the marginal probability of an FGS of family i. Assuming different families and different sample clusters are genetically independent given their matching configurations, the conditional probability of all FGSs $\{\mathbf{f}_j\}$ and clusters of SFSs $\{\mathbf{c}_j\}$, given the matching matrix $\mathbf{z}$, is:

$$\begin{aligned} p(\{\mathbf{c}_j\}, \{\mathbf{f}_i\} \mid \mathbf{z}) &= \prod_j p(\mathbf{c}_j \mid \{\mathbf{f}_i\}, \mathbf{z}) \prod_i p(\mathbf{f}_i) \\ &= \prod_{ij} \pi(\mathbf{c}_j \mid \mathbf{f}_i)^{z_{ij}} \prod_j \pi(\mathbf{c}_j)^{1-\Sigma_i z_{ij}} \prod_i p(\mathbf{f}_i) \\ &= \prod_{ij} \pi(\mathbf{c}_j \mid \mathbf{f}_i)^{z_{ij}} \prod_i p(\mathbf{f}_i). \end{aligned}$$

Note that according to the constraints of one-to-one matching in Eq. (2), we have $1 - \Sigma_i z_{ij} = 0$.

The likelihood ratio of an overall matching specification $\mathbf{z}$ versus a null hypothesis (that all families and samples are unrelated) is:

$$\begin{aligned} LR(\mathbf{z}) &= \frac{p(\{\mathbf{c}_j\}, \{\mathbf{f}_i\} \mid \mathbf{z})}{p(\{\mathbf{c}_j\}) p(\{\mathbf{f}_i\})} \\ &= \frac{\prod_j \prod_i \pi(\mathbf{c}_j \mid \mathbf{f}_i)^{z_{ij}}}{\prod_j \pi(\mathbf{c}_j)} \\ &= \prod_{ij} \left[\frac{\pi(\mathbf{c}_j \mid \mathbf{f}_i)}{\pi(\mathbf{c}_j)} \right]^{z_{ij}}. \end{aligned} \tag{3}$$

Let $\Lambda_{ij} \equiv \pi(\mathbf{c}_j \mid \mathbf{f}_i)/\pi(\mathbf{c}_j)$, and take the logarithm of LR, we have

$$\log LR(\mathbf{z}) = \sum_{j=1}^{J} \sum_{i=1}^{N} z_{ij} \log \Lambda_{ij}. \tag{4}$$

We postulate that an optimal body identification corresponds to a $\mathbf{z}$ that maximizes the likelihood ratio of matching family-clusters versus randomly generated $\{\mathbf{c}_j\}$ and $\{\mathbf{f}_i\}$. In the sequel we describe algorithms for identifying the sample clusters from the SGSs of samples, and for solving the optimal matching.

3.2 Sample clustering

The first problem in body identification is to determine the total number of victims in the case, and group body remains for each victim. We determine whether two samples, $\mathbf{s}_i$ and $\mathbf{s}_j$, are from the same victim or not based on the ratio of their joint probabilities under the two circumstances:

$$LR(\mathbf{s}_i, \mathbf{s}_j) = \frac{p(\mathbf{s}_i, \mathbf{s}_j)}{p(\mathbf{s}_i) p(\mathbf{s}_j)} = \frac{p(\mathbf{s}_i \mid \mathbf{s}_j)}{p(\mathbf{s}_i)} = \prod_{k=1}^{K} \frac{p(s_{ik} \mid s_{jk})}{p(s_{ik})}$$

The conditional probability $p(s_{ik} \mid s_{jk})$ of genotypes will be referred to as an *error model*, which will be specified in §3.2.2.

3.2.1 The union-find clustering algorithm Let each sample in the case be represented by a node, we can define an undirected graph over all samples of interest. Two nodes are connected if $LR(\mathbf{s}_i, \mathbf{s}_j) > \theta_c$, where θ_c is a user-specifiable threshold. As a common practice in mass disaster forensic identification, any two samples with more than two genotypes differences are immediately considered disconnected. Sample clustering is done by partition this graph into connected subgraph, which can be implemented efficiently using a *union-find algorithm*. We defines three operations: **make-set**—creates a set, **union**—merges two sets, and **find**—returns the host set of a node. The algorithm proceeds as follows:

(1) **make-set** creates a set for each node

(2) For two nodes of each edge, iterate the following

- **find** the corresponding sets,
- **union** the two sets (if they are connected by cross-set edges).

This process will converge to a clustering of samples, without a prior specification the number of clusters, but a threshold controlling the tightness of the clusters. This is a desirable feature in forensic inference because usually the legal agents would need to leverage their forensic experience and determine tolerable risk of legal decisions circumstantially. Once the clustering is complete, we extract a consensus SGS $\hat{\mathbf{c}}_j$ for each cluster $\mathbf{c}_j$ based on a maximum likelihood principle. That is, given the consensus $\hat{\mathbf{c}}_j$ that corresponds to the true genetic state (TGS) of a victim, the conditional probability of all SGSs of this cluster (i.e., this victim) is maximized:

$$\begin{aligned} \hat{\mathbf{c}}_j &= \arg \max_{\mathbf{t}} p(\mathbf{t}) \prod_{l \in \mathbf{c}_i} p(\mathbf{s}_l \mid \mathbf{t}) \\ &= \arg \max_{\mathbf{t}} \prod_{k=1}^{K} \Big(p(t_k) \prod_{l \in \mathbf{c}_j} p(s_{lk} \mid t_k) \Big), \end{aligned}$$

where the marker-specific conditional probability $p(s_{lk} \mid t_k)$ is given by the error model described bellow.

3.2.2 The error model The error model defines the probability distribution of a marker-specific sample genotype given the true genotype, $p(s_k \mid t_k)$. For two alleles $a \neq b$ of any markers (i.e., locus), we define five error types:

(1) Measurement error: Allele a is misread as $a \pm 0.1$ by the technician

(2) Calibration error: True genotype is (a, b) but calibration ladder is off by one, so instruments shows $(a+1, b+1)$ or $(a-1, b-1)$

(3) PCR Shutter error: True genotype is (a, a) but instruments shows $(a, a \pm 1)$

(4) Threshold error: True genotype is (a, b) but the b signal falls below threshold, so instruments shows (a, a)

(5) Mutation error: Allele a mutates to allele b

The probability of measurement, calibration, shutter, and threshold error are constants, denoted as ϵ_m, ϵ_c, ϵ_s, ϵ_t respectively. Based on the stepwise mutational model (Valdes *et al*., 1993) for microsatellite, the probability of a mutation from a to b is $p(b \mid a) = 0.5\mu(1 - \alpha)\alpha^{|b-a|-1}$, where μ is the mutation rate (probability of any mutation) and α is the factor by which mutation decreases as distance increases. Although this mutation distribution is not stationary (i.e. it does not ensure allele frequencies to be constant over the generations), it is simple and commonly used in forensic inference. Shutter, threshold, and calibration errors are defined on genotypes, but measurement and mutation errors are defined on alleles and have to consider two combinations, $p(s_k^0 \mid t_k^0)p(s_k^1 \mid t_k^1)$ and $p(s_k^0 \mid t_k^1)p(s_k^1 \mid t_k^0)$. To summarize, for $s_k \neq t_k$, we have:

$$p(s_k \mid t_k) = \begin{cases} \epsilon_c & \text{if } s_k^0 - t_k^0 = s_k^1 - t_k^1 = \pm 1 \\ \epsilon_s & \text{if } s_k^0 = s_k^1 = t_k^0,\ |s_k^1 - t_k^1| = 1 \\ \epsilon_t & \text{if } s_k^0 = t_k^0 = t_k^1 \\ \max(q(s_k^0; t_k^0)q(s_k^1; t_k^1), q(s_k^0; t_k^1)q(s_k^1; t_k^0)) & \text{otherwise} \end{cases},$$

where the allele error function $q(b; a)$ is defined as

$$q(b; a) = \begin{cases} 1 & \text{if } b = a \\ \epsilon_m & \text{if } |b - a| = 0.1 \\ 0.5\mu(1 - \alpha)\alpha^{|b-a|-1} & \text{otherwise} \end{cases}.$$

The $p(s_k \mid t_k)$ is a conditional probability that must sum to one. Thus, we define the "consistence" probability $p(s_k = t_k \mid t_k)$ as one minus all error probabilities, which is large comparing to the overall error probability (since the probabilities of each error type are always set to be very small):

$$p(s_k = t_k \mid t_k) = 1 - \sum_{s_k \neq t_k} p(s_k \mid t_k).$$

3.3 Pedigree inference

The conditional probability of a TGS given the FGS of a matching family, $p(\hat{\mathbf{c}}_j \mid \mathbf{f}_i)$, can be derived by pedigree inference. As discussed in Lauritzen and Sheehan (2003), the joint distribution of $\{\hat{\mathbf{c}}_j, \mathbf{f}_i\}$ defined by an arbitrary pedigree can be specified by a *probabilistic graphical model* (Pearl, 1988; Cowell *et al*., 1999), or more specifically, a *Bayesian network* (Pearl, 1986).

Recall that an FGS $\mathbf{f}_i$ is a two-tuple of donor genotypes $\{\mathbf{d}_1, \ldots, \mathbf{d}_{n_i}\}$ and a familial pedigree T_i. Based on T_i, we can construct a particular Bayesian network, known as *allele network*, or *gene pedigree* (Lauritzen and Sheehan, 2003), for all the alleles from all members (donor and non-donor) of the family and from the purported victim. Assuming that markers are independent and following the same pedigree, we construct an allele network for a single marker, say microsatellite k, as follows. For each individual, we introduce two allelic nodes, u_k^0 and u_k^1 (which are unobserved), denoting the maternal and paternal allele of this individual, respectively; and a genotype node u_k^g, which are observed for the donors and hidden for the non-donors in the family. Since the genotype is determined jointly by the two alleles, we have arcs pointing from each allelic node to its corresponding genotype node (Fig. 1 and Fig. 2). Due to Mendelian inheritance, the marker alleles in a decedent is dependent on that in his/her direct parents, thus we also have arcs pointing from the allelic nodes of a parent to the allelic nodes of the children. Note that the allelic nodes of individuals that are *founder* of the pedigree do not have any arcs pointing to them. For those individuals who are donors in a family (i.e., their genotype states are available from their DNA samples), we denote their corresponding genotype nodes as observed variables, shown as shaded circles. The genotype of the purported victim is also observed via sample clustering, but need to be matched correctly. In Fig. 1 and Fig. 2 we use circles with thick border to denote the genotype of a *candidate* victim. Because markers are independent in our case, each marker has a separate allele network with the same structure but different donor evidence (i.e., marker-specific genotypes). The joint probability of multiple markers is the product of all locus-specific marker probabilities defined by the allele network. Specifically, we use the following conditional distributions in our allele network model:

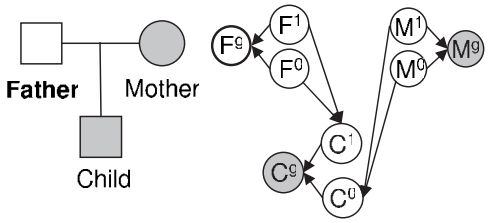

Fig. 1. A simple pedigree and its allele network, shaded nodes as donors and bold nodes as victim.

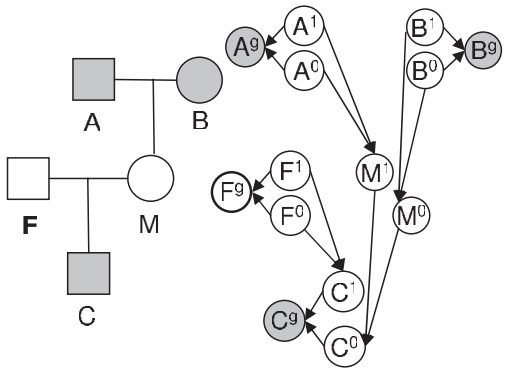

Fig. 2. A pedigree of three generations and its allele network.

(1) Founder distribution: $p(u_k^e) = \pi_{k,\,u_k^e}$, where $e \in \{0, 1\}$ represents the parental index of the allele, $\pi_{k,a}$ is the population frequency of allele a.

(2) Meiosis distribution: For an allele t_k^e inherited from a parent with genotype $s_k = \{u_k^0, u_k^1\}$, we have

$$p(t_k^e \mid u_k^0, u_k^1) = \begin{cases} 0.5 & \text{if } t_k^e = u_k^0 \text{ or } t_k^e = u_k^1, \text{ and } u_k^0 \neq u_k^1, \\ 1 & \text{if } t_k^e = u_k^0, \text{ and } u_k^0 = u_k^1, \\ 0 & \text{otherwise.} \end{cases}$$

(3) Genotype distribution: $p(u_k^g \mid u_k^0, u_k^1)$, which is specified by the error model defined in §3.2.2.

Given the allele network, and the above conditional distributions of a node in the network given its graph parents (not to be confused with biological parents), one can write down the joint distribution of all nodes, i.e. the victim and the FGS, as a product of all node-specific conditionals following a natural node ordering (e.g., from founder to decedents) (Pearl, 1988). From this joint probability we can derive conditional probability $p(x_F \mid x_E)$ of a set of variables $F \subsetneq V$ conditioned on a set of observed variables $E \subsetneq V$. F is called

query nodes, E is called *evidence nodes* and V is the totality of all nodes. The junction tree algorithm (Lauritzen and Spiegelhalter, 1988) can perform exact inference efficiently on a network of reasonable size, which is sufficient for our purpose.

3.4 Viterbi match: optimal body identification via linear programming

Given the conditional probabilities of TGSs of sample clusters and the FGSs of their matching families, $p(\hat{\mathbf{c}}_j \mid \mathbf{f}_i)$, now we are ready to tackle the optimal matching between sample clusters and families. Let us view the match matrix $\mathbf{z}$ as a representation of the edge configuration of a bipartite graph in which the clusters $\{\hat{\mathbf{c}}_j\}$ correspond to nodes in one partite, and the families $\{\hat{\mathbf{c}}_j\}$ correspond to the nodes in the other partite. Associating each edge between $\{\hat{\mathbf{c}}_j\}$ and $\mathbf{f}_i$ with a weight equal to $\log \pi(\hat{\mathbf{c}}_j \mid \mathbf{f}_i)/\pi(\hat{\mathbf{c}}_j)$, then the total cost of the matching, $LR(\mathbf{z})$, corresponds to the sum of weights of edges in the bipartite graph. Finding an optimal matching is equivalent to the classical maximum weight bipartite matching problem. We can solve this bipartite matching problem by mixed integer linear programming (LP):

$$\begin{aligned} &\max \sum_{j=1}^{J}\sum_{i=1}^{N} z_{ij} \log \Lambda_{ij} \\ &z_{ij} \in \{0,1\},\ \sum_{i=1}^{N} z_{ij} = 1\ \forall j,\ \sum_{j=1}^{J} z_{ij} = 1\ \forall i. \end{aligned} \tag{5}$$

There are many efficient algorithms and implementation for solving the above LP, and we use the open source Gnu Linear Programming Kit (GLPK) (Makhorin, 2001). Note that this approach gives a globally optimal mapping assignment between (equal number of) clusters and samples, analogous to finding the Viterbi path in hidden Markov model (but in this case an optimal matrix). Thus, we call the resulting body identification results a *Viterbi match*.

4 POSTERIOR MATCH AND MATCHING DISAMBIGUATION

The one-to-one constrain assumed so far in our algorithm is not always valid. In fact, since we cluster samples based on a tightness threshold rather than a given fixed number of clusters, we can not easily enforce $N = J$. In practice, a cluster may be unmatched, i.e. not assigned to any reporting family (e.g., due to poor sample quality, or nonexistence of the true claiming family); conversely, a family may also be unmatched (e.g., because no remain of the victim is found).

We assume each sample either comes from one family, or it is a random sample from the population. However, samples from one victim may be clustered into multiple clusters due to heterogeneity of the physical and measurement quality of different samples. To accommodate these flexibilities, we relax the normality constraints on the columns and rows of matching matrix $\mathbf{z}$, so that multiple clusters can be matched to one family, or no clusters or family get matched:

$$\sum_{i=1}^{N} z_{ij} \in \{0,1\}\ \forall j. \tag{6}$$

Furthermore, instead of seeking an overall estimate of $\mathbf{z}$, we would like to have a confidence measure of each of the judgments (i.e., match or not-match) specified by $\mathbf{z}$. From a forensic perspective, only matches with small enough probability should be considered (forensically) impossible, and excluded from legal consideration. In the sequel, we show how to calculate the posterior probability of a matching given cluster and family data; and then we show that, with this probability, how to screen out unambiguous matching outcomes and extract the truly ambiguous cases that merit costly personalized forensic investigation.

4.1 Posterior probability of a many-to-one matching

Now we derive the posterior probability of a matching given cluster TGSs and family FGSs, $p(\mathbf{z} \mid \{\mathbf{c}_j\}, \{\mathbf{f}_i\})$. According to the Bayes' theorem, we have:

$$p(\mathbf{z} \mid \{\mathbf{c}_j\}, \{\mathbf{f}_i\}) = \frac{p(\mathbf{z}) p(\{\mathbf{c}_j\}, \{\mathbf{f}_i\} \mid \mathbf{z})}{p(\{\mathbf{c}_j\}, \{\mathbf{f}_i\})}. \tag{7}$$

Since we do not know the matching *a priori*, $p(\mathbf{z})$ can be taken as uniform. Following the notations in §3.1, let $p(\mathbf{f}_i)$ and $\pi(\hat{\mathbf{c}}_j)$ denote the marginal probabilities of a given family, and a cluster TGS, respectively; and let $\pi(\hat{\mathbf{c}}_j \mid \mathbf{f}_i)$ denote the conditional probability a cluster TGS $\hat{\mathbf{c}}_j$ given its matching FGS $\mathbf{f}_i$ (i.e., $z_{ij} = 1$). Following the new constrain given by Eq. (6), and since the cluster TGSs are independent of each other given a matching $\mathbf{z}$, the conditional probability of each cluster TGS given a matching is:

$$p(\hat{\mathbf{c}}_j \mid \{\mathbf{f}_i\}, \mathbf{z}) = \begin{cases} \pi(\hat{\mathbf{c}}_j \mid \mathbf{f}_i) & \text{if } \exists i : z_{ij} = 1 \\ \pi(\hat{\mathbf{c}}_j) & \text{if } \sum_l z_{lj} = 0 \end{cases}, \tag{8}$$

Therefore the joint conditional probability of the TGSs and FGSs given $\mathbf{z}$ is

$$\begin{aligned} & p(\{\hat{\mathbf{c}}_j\}, \{\mathbf{f}_i\} \mid \mathbf{z}) \\ = \ & p(\{\hat{\mathbf{c}}_j\} \mid \{\mathbf{f}_i\}, \mathbf{z}) p(\{\mathbf{f}_i\} \mid \mathbf{z}) \\ = \ & \prod_j p(\hat{\mathbf{c}}_j \mid \{\mathbf{f}_i\}, \mathbf{z}) \prod_i p(\mathbf{f}_i) \\ = \ & \prod_{ij} \pi(\hat{\mathbf{c}}_j \mid \mathbf{f}_i)^{z_{ij}} \prod_j \pi(\hat{\mathbf{c}}_j)^{1-\sum_l z_{lj}} \prod_i p(\mathbf{f}_i) \\ = \ & \prod_{ij} \left[\frac{\pi(\hat{\mathbf{c}}_j \mid \mathbf{f}_i)}{\pi(\hat{\mathbf{c}}_j)}\right]^{z_{ij}} \prod_j \pi(\hat{\mathbf{c}}_j) \prod_i p(\mathbf{f}_i) \\ = \ & \prod_{ij} \Lambda_{ij}^{z_{ij}} \prod_j \pi(\hat{\mathbf{c}}_j) \prod_i p(\mathbf{f}_i). \end{aligned}$$

Thus, Eq. (7) reduces to:

$$p(\mathbf{z} \mid \{\mathbf{c}_j\}, \{\mathbf{f}_i\}) = \frac{1}{A} \prod_{ij} \Lambda_{ij}^{z_{ij}}, \tag{9}$$

where A is a normalizing constant summing over all $\mathbf{z}$. Using the fact that we are summing over all possible $\mathbf{z}$ under limitation (6), we can derive normalizing constant in closed form:

$$A = \sum_{\mathbf{z}} \prod_j \prod_i \Lambda_{ij}^{z_{ij}} = \prod_j (1 + \sum_i \Lambda_{ij}). \tag{10}$$

According to Eqs. (10) and (9), now we have a close-form expression for the posterior probability of a matching given the clusters and families data:

$$p(\mathbf{z} \mid \{\mathbf{c}_j\}, \{\mathbf{f}_i\}) = \frac{\prod_{ij} \Lambda_{ij}^{z_{ij}}}{\prod_j (1 + \sum_i \Lambda_{ij})}. \tag{11}$$

4.2 Individual posterior match and matching disambiguation

To qualify a candidate match, $\mathbf{c}_j$ versus $\mathbf{f}_i$, we compute the posterior probability of a match as follows. Let $\mathbb{Z}_{ij}$ denote the set of all matrix $\mathbf{z}$ in which $z_{ij} = 1$, i.e. all possible matching that assigns $\mathbf{c}_j$ to $\mathbf{f}_i$:

$$\mathbb{Z}_{ij} = \{\mathbf{z} : z_{ij} = 1\}, \tag{12}$$

Similarly, let $\mathbb{Z}_{ij}^c$ denote the complement of this set. Now the posterior probability of an *individual posterior match* (IPM) given TFSs of all samples clusters and FGSs of all reporting families can be computed as:

$$p(z_{ij} = 1 \mid \{\mathbf{c}_m\}, \{\mathbf{f}_l\}) = \sum_{z \in \mathbb{Z}_{ij}} p(\mathbf{z} \mid \{\mathbf{c}_m\}, \{\mathbf{f}_l\}) \tag{13}$$

To disqualify a candidate pair, $\mathbf{c}_j$ and $\mathbf{f}_i$, on the basis that they are extremely unlikely to be a true match, we define our *decoupling confidence* (DC) of this pair to be the posterior probability mass of the set $\mathbb{Z}_{ij}^c$, which can be computed as follows:

$$\begin{aligned}
& p(\mathbf{z} \in \mathbb{Z}_{ij}^c \mid \{\mathbf{c}_m\}, \{\mathbf{f}_l\}) \\
= \;& 1 - p(\mathbf{z} \in \mathbb{Z}_{ij} \mid \{\mathbf{c}_m\}, \{\mathbf{f}_l\}) \\
= \;& 1 - \sum_{\mathbf{z} \in \mathbb{Z}_{ij}} p(\mathbf{z} \mid \{\mathbf{c}_m\}, \{\mathbf{f}_l\}) \\
= \;& 1 - \sum_{\mathbf{z} \in \mathbb{Z}_{ij}} \tfrac{1}{A} \prod_m \prod_l \Lambda_{lm}^{z_{lm}} \\
= \;& 1 - \tfrac{1}{A} \Lambda_{ij} \prod_{m \neq j} \Big(1 + \sum_l \Lambda_{lm}\Big) \\
= \;& 1 - \frac{\Lambda_{ij} \prod_{m \neq j} \Big(1 + \sum_l \Lambda_{lm}\Big)}{\prod_m \Big(1 + \sum_l \Lambda_{lm}\Big)} \\
= \;& 1 - \frac{\Lambda_{ij}}{1 + \sum_l \Lambda_{lm}}.
\end{aligned}$$

Given the posterior probabilities of all IPMs, and the values of all DCs, now we can not only extract *maximum a posterior* (MAP) matches as in § 3, but also perform a *matching disambiguation* for the given $\{\mathbf{c}_m\}$ and $\{\mathbf{f}_l\}$. Essentially, for the later task we exclude a candidate match with DC higher than a specifiable threshold $1 - \theta_m$. Different values can be assigned to θ_m based on the situation of the disaster, and $\theta_m = 10^{-6}$ is commonly used in mass disaster scenes, meaning that by excluding the chosen pair of cluster TGS and family FGS, in less than one out of a million cases we missed a true match. If the DCs of all family-cluster pairs are higher than $1 - \theta_m$, then we are confident the cluster is unmatched, i.e. no family claims this victim.

After the aforementioned impossible-match exclusion, if there is zero or only one possible family for a cluster, this cluster is unambiguous and is considered determined. Otherwise, if a remaining cluster-family pair passes an IPM threshold, it is still considered a valid match. Finally, the clusters that still have ambiguity, i.e., with two or more possible families of IPM lower than the threshold, will be reported to human expert for further forensic investigate.

5 EXPERIMENTS

Due to high confidentiality of forensic DNA fingerprint data, a common practice in forensic science is to validate the models and algorithms via computer simulation experiments, for which the true matchings are known. Following convention, thirteen FBI CODIS markers are used. In each experiment we simulate N core families from a single population, by generating two random parents based on population allele frequencies, and generating one child from the parents. The victim is the child in three simulations, and in two other simulations the victim is one of the parents. Allele frequencies $\pi_{k,a}$ are assumed to be known and correct. Then we generate several TGSs for each victim, using the error model with different values of the parameters (to simulate different level of noise). The number of SGSs generated from a victim is distributed uniformly in an interval, $[M^{(0)}, M^{(1)}]$. Throughout the experiments, the parameters used for sample generation are intentionally set to be different from the ones used in our later inference, so that our test is unbiased and objective. For each marker, there is a probability of ϵ_u that the genotype is missing. The simulating parameter ϵ_u is set to be high, to represent that some samples are heavily degraded. However we require that the total number of available markers to be greater than 4 to make our cases forensically realistic—for situations where the recovered markers are less than or equal to 4, DNA evidence are usually dismissed due to lack of reliability. We performed five experiments with different simulating parameters, as described below:

(1) $N = 100$, $[M^{(0)}, M^{(1)}] = [3,7]$, so on average 500 samples. Victim is the child, and donors are the two parents. Simulation parameters are $\epsilon_u = 1/10$, $\epsilon_m = \epsilon_c = 0.001$, $\epsilon_s = \epsilon_t = 0.004$.

(2) A noisier setting, $N = 100$, $[M^{(0)}, M^{(1)}] = [3, 7]$, so on average 500 samples. Victim is one of the parents, and donors are the child and the other parent. Simulation parameters are $\epsilon_u = 1/4$, $\epsilon_m = \epsilon_c = 0.001$, $\epsilon_s = \epsilon_t = 0.004$.

(3) Similar to simulation 2 but with even more noise: $N = 100$, $[M^{(0)}, M^{(1)}] = [1, 9]$, so on average still 500 samples, but the cluster sizes vary more. The values of the simulation parameters are now higher, $\epsilon_u = 1/3$, $\epsilon_m = \epsilon_c = 0.002$, $\epsilon_s = \epsilon_t = 0.008$.

(4) Similar to simulation 1 but contains 500 families and on average 2500 samples (1,250,000 potential matches).

(5) Similar to simulation 1 but contains 1000 families and on average 5000 samples (5,000,000 potential matches).

The parameters used during computational inference in all four experiments are the same: $\epsilon_m = 0.00025$, $\epsilon_c = 0.00025$, $\epsilon_s = 0.001$, $\epsilon_t = 0.001$, which may be different from the parameters for sample simulation. The clustering LR threshold is $\theta_c = 500$. All experiments are repeated 9 times and their results are averaged.

5.1 Results on optimal body identification

Since our clustering is stringent, the number of resulting clusters is always greater or equal to the number of families ($N \leq J$), and the assumption of one-to-one mapping behind the Viterbi matching via LP no longer holds. We can still apply LP by enforcing the same optimization and constraint terms in Eq. (5), which means we still require one matching family for each cluster and one matching cluster for each family, but some clusters may be unmatched.

We perform optimal body identification using Viterbi matching via LP and MAP matching. We measure the performance by average false-negative rate (FN) and false-positive rate (FP), where FN is the ratio of undiscovered true matches to all true matches, and FP

Table 1. Optimal body identification performance of LP and MAP

Sim	LP FN	LP FP	MAP FN	MAP FP
1	0.0109	0.0	0.0	0.0
2	0.0130	0.0	0.0043	0.0043
3	0.0567	0.0112	0.0225	0.0225
4	0.0099	0.0004	0.0020	0.0020
5	0.0073	0.0002	0.0021	0.0021

Comparison of average false-negative (FN) and false-positive (FP) rate of LP and MAP algorithm. LP denotes the Viterbi match via LP based on one-to-one mapping assumption in § 3.4, and MAP denotes the MAP match based on many-to-one mapping in § 4.2.

is the ratio of incorrect predictions to all predictions. The results are shown in Table 1.

Overall, LP has low FP, but the FN is very high, mainly due to incorrectness of the one-to-one assumption in the model. MAP has slightly higher FP, but the FN is much lower. In simulation 1, MAP has zero FN and FP. Overall, both algorithms have good performance, even in the presence of noise and incomplete information. We are not aware of existence of any algorithm or software for this kind of forensic task in earlier and current literature.

5.2 Results on matching disambiguation

In a matching disambiguation task, our goal is to reduce as much as possible the amount of human effort in forensic inference by remove impossible cluster-family matches and high-confidence matches from a given mass disaster case. In this section, we compare the disambiguation results using the individual posterior match method with the ones using a conventional approach that excludes a candidate match by thresholding the likelihood ratio, e.g., a candidate match from $\mathbf{c}_j$ to $\mathbf{f}_i$ is excluded (i.e., deemed impossible) if $\Lambda_{ij} < \theta_m = 10^{-6}$. Such threshold means that the relative probability of a cluster-family match is only 10^{-6} compared to an alliterative hypothesis that they are unrelated.

We found that the accuracy of disambiguation via the posterior methods is significantly better than that of the conventional LR thresholding approach, as shown in Table 2. The threshold θ_m is set to be 10^{-6} in both algorithms. In our experiments, the accuracies are measured by: (1) the average percentage of remaining ambiguous clusters; (2) the average percentage of remaining ambiguous matching families for each cluster; and (3) the ratio of ambiguous family-cluster matches over all candidate matches. After applying the posterior match disambiguation algorithm, the remaining ambiguous clusters are almost always single samples. On average, the 500 samples were reduced to only 1, 5, and 13 ambiguous samples, in simulation 1, 2, and 3, respectively; and each ambiguous cluster has 6, 8, and 10 ambiguous candidate matching families, respectively. In simulation 4, 2500 samples and 500 families were reduced to 5 samples, each having 21 candidate families. In simulation 5, 5000 samples and 1000 families were reduced to 6 samples, each having 33 candidate families. Under the same noise level, larger sample size results in better reduction rate. The results of LR thresholding is generally much worse, about 3 to 12 fold increase in cluster ambiguity, and 3 to 5 fold increase in overall ambiguity.

Table 2. Comparison of disambiguation by posterior threshold and by LR threshold

Sim	Posterior Clusters	Posterior Families	Posterior Matches	LR thresholding Clusters	LR thresholding Families	LR thresholding Matches
1	0.01	0.06	0.0007	0.03	0.07	0.0019
2	0.04	0.08	0.0034	0.48	0.04	0.0190
3	0.12	0.10	0.0119	0.53	0.07	0.0371
4	0.01	0.04	0.0004	0.08	0.02	0.0013
5	0.01	0.03	0.0002	0.14	0.01	0.0010

Results of disambiguation by posterior and LR threshold. ''Clusters'' denote the average percentage of remaining ambiguous clusters. ''Families'' denote the percentage of ambiguous candidate matching families for each of these clusters. ''Matches'' denotes the ratio of ambiguous family-cluster matches over all possible matches. Parameter settings of the three simulations are described in 5.

A close examination of our results showed that these ambiguities all occurred in samples with severely degraded markers, typically with only 5 of the 13 marker readable. Under these circumstances, a family becomes a candidate match to a sample even when only 3 of the markers are compatible with that of the samples within an error range. In practice, such genetic samples would automatically be ruled legally insubstantiative even before computational forensic inference is conducted, and would require additional forensic evidence. Thus, our disambiguation results presented above is in fact a worst-case result, and the actual rate of disambiguation in real life can be much better if we are willing to insist on more stringent requirement for the quality of the DNA samples (e.g., by requiring more than half of the markers can be clearly typed). It is noteworthy that a domain expert does not need to examine the ambiguous families of each cluster one by one. An expert can determine the true family from evidences other than DNA, or determine the sample as unidentifiable, or repeat the DNA sampling.

5.3 Analysis of disambiguation threshold

The major difference between the posterior disambiguation and the LR-based method is that posterior disambiguation relates the LRs of all possible families versus a candidate cluster when inferring about each single matching. That is, for one cluster, if several likely matching families already exist, other families with lower LRs will be considered less likely, whereas in the conventional LR-based disambiguation, each candidate matching is assessed independent of other candidates. We illustrate this difference in disambiguation criteria in Figure 3. The histogram of all the log LR of simulation 1 and 2 is shown in Figure 3A and 3C. For the log LR of all possible families corresponding to a well-typed (i.e., with most markers measurable) cluster, as shown in Figure 3B and 3E, usually there are only a few (in this case, only one) candidate matches having LR above 10^{-6}, so the two methods make little (or no) difference because of nearly inexistence of between-match influences. However, for a degraded cluster illustrated in Figure 3C and 3F, there are many candidate matches with large LRs and they influence each other. Consequently the disambiguation via posterior inference tends to assess other candidates to be less likely than would have been suggested by the LRs alone. This effectively results in a criterion more stringent than 10^{-6}. The LR thresholding approach, on the other hand, still use the same threshold on LR. As shown in Figure 3C and 3F, the posterior match

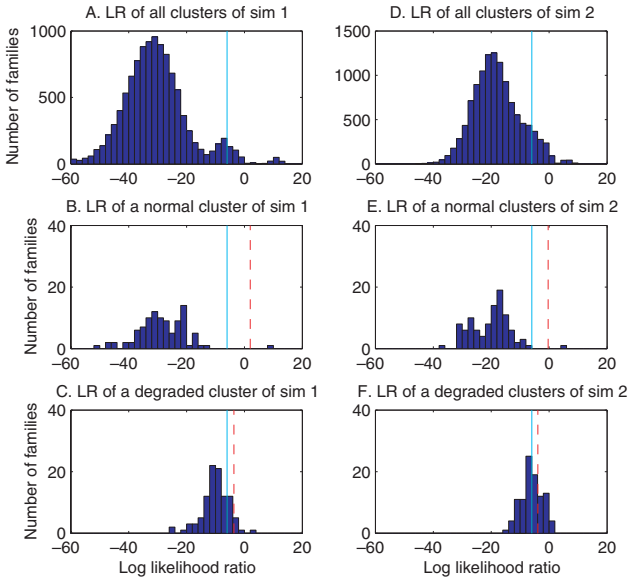

Fig. 3. The histogram of log likelihood ratio of simulation 1 and 2. **A–C** is based on simulation 1 and **D–F** is based on simulation 2. The x-axis is common logarithm of likelihood ratio, and the y-axis is number of families or matches. Vertical blue solid line denotes 10^{-6} threshold, and red dotted line denotes the effective threshold of disambiguation corresponding to the posterior match criteria. Specifically, we have: **A**. Distribution of all sample clusters of simulation 1. **B**. LR distribution of a well-typed cluster of simulation 1. **C**. LR distribution of a degraded cluster of simulation 1. **D–F**. The LR distributions of all sample clusters, a normal cluster, and a degraded cluster, respectively, in simulation 2.

method can reduces the ambiguity by a half or even more for degraded clusters.

In traditional forensic identification cases, which do not deal with DNA sample clustering but consider mostly high-quality anonymous samples, the LR of the correct identification tends to be very high, and there is usually no ambiguity. To see the difference in a mass disaster case, it is instructive to take a close look at the dataset and the ambiguous clusters and families reported by our algorithm. When there are fewer than 7 markers in a sample, typically there are indeed many ambiguous family pedigrees that cannot be excluded from a forensic perspective. For example, consider the highly degraded samples, of which an example is shown in Table 3. Typically such samples can have multiples plausible matching families, and the matches listed in Table 3 are only a few of all the likely matches. The ambiguity problem become very serious when the quality of the samples gets really poor, e.g., with fewer than 5 usable markers available. Essentially, the evidence become not enough for body identification—given only three or four markers, there could be too many perfect matches. In this case, the power of any computational and/or manual forensic inference diminishes, and we must seek additional evidence. We discuss some of the options in the next section.

6 DISCUSSION

Extending our probabilistic forensic inference methods to include other evidence is straightforward. For example, sometimes, in the forensic samples there also exist sequence data from the two segments of the hyper-variable control regions (e.g., regions 16,024 to 16,365 and 73 to 340) of the 16,569bp human mitochondria DNA (mtDNA). Because mtDNA has far more copies than the genome, they are often sequenceable when the genome is degraded and not sequenceable. Inheritance of mtDNA is maternal only, so there is much less uncertainty. But the mtDNA is less variable compared to microsatellites in genomic DNA. For example, while there are in principle 10 or more possible SNP differences in the mtDNA between any two individuals, a match is not conclusive due to high degeneracy of these polymorphism in human population. For example, about 7% of all Caucasian males have the same mtDNA sequence. Nevertheless, mtDNA can still be used to eliminate impossible matches, i.e., we can remove cluster-family matches with inconsistent mtDNA, and further reduce ambiguity.

Table 3. Case study of a highly degraded sample

Errors	Log LR	Description	THO1	D7S820	VWA
		Sample	(7,8)	(8,11)	(14,15)
0	1.70	True mate	(6,9)	(10,11)	(13,15)
		True child	(8,9)	(8,10)	(13,15)
0	1.00	Mate	(6,7)	(10,11)	(15,17)
		Child	(7,8)	(11,11)	(15,17)
1	−1.66	Mate	(7,9)	(9,10)	(18,18)
		Child	(6^{μ},9)	(10,11)	(15,18)
2	−4.12	Mate	(9,9.3)	(8,11)	(17,18)
		Child	(7,9)	($8,9^{s}$)	($18,18^{t}$)

A highly degraded sample of which three typed markers are shown. THO1, D7S820, and VWA are three markers in the CODIS system. The symbols a_s, a_t, a_{mn} denotes shutter, threshold, mutation error respectively. All the pedigrees have one of the parents as the victim and the other parent and a child as the donors. Among candidate families with high LR, four representative matches are listed here. Note that many different combinations are qualified for a match.

Occasionally, there will also be alleged direct sample evidence for a victim from a personal effect, such as a comb or tooth brush, in which case the genotype is available for the victim in the relevant family pedigree. Similarly, other factors like gender and blood type can be easily included using probabilistic rules.

In mass disaster scenes it is important to validate pedigree structure and donor evidence. For example, there may be an error in some donor's genotype, making it inconsistent with other donors' genotype. There is also the rather delicate issue that sometimes paternity or other blood relationships are not true. This kind of error can be detected by calculating the marginal probability of the evidence based on the allele network model. Families with probabilities under a threshold can be picked out and given to experts for examination. A family may have several victims in a mass disaster site. In this case one can introduce duplicated pedigrees one for each alleged victims. Each pedigree has the same structure and donor genotypes, but has different victim node. One must be careful about now the incorrectness of independence assumption for all pedigrees and for all the victim samples. For example, if a father and his son are both victims, their genotypes are not independent. This could slightly complicate the probabilistic inference computation for LR-based Viterbi match and posterior match.

Finally, it is noteworthy that, although in current forensic applications, genetic markers are usually chosen as independent

(e.g. the thirteen CODIS markers reside on different chromosomes), our probabilistic framework presented in this paper does not rely on the assumption that markers are independent. In extremely degraded disaster scenes, using single nucleotide polymorphism (SNP) for identification may be helpful (Cash *et al*., 2003); and for SNPs with high linkage disequilibrium, the markers are no longer independent. In such cases we can create an allele network with linkage probability, by adding a meiosis variable which couples different markers (Lauritzen and Sheehan, 2003). Under such circumstances, the allele network will become more complex and approximate inference or sampling may be necessary (Jordan *et al*., 1999; Xing *et al*., 2003).

In conclusion, we have presented a probabilistic modeling and inference framework for mass disaster victim identification. We expect that this framework can be easily generalized to handle more complicated forensic inference problems, and leverage richer forensic evidence or expert knowledge. It offers a promising platform to develop automatic expert system for a wide-range of forensic and genetic inference applications.

REFERENCES

Ballantyne,J. (1997) Mass disaster genetics. *Natural Genetics*, **15**, 329–331.

Cash,D.C., Hoyle,J.W. and Sutton,A.J. (2003) Development under extreme conditions: forensic bioinformatics in the wake of the World Trade Center disaster. In *Proceedings of Pacific Symposium on Biocomputing 2003,* **8**, 638–653.

Cowell,R.G., Dawed,A.P., Lauritzen,S.L. and Spiegelhalter,D.J. (1999) *Probabilistic Networks and Expert Systems*. Springer, New York.

Egeland,T., Mostad,P.F., Stenersen,M. and Mevag,B. (2000) Beyond traditional paternity and identification cases: selecting the most probable pedigree. *Forensic Science International*, **110**, 47–59.

Hsu,C.M., Huang,N.E., Tsai,L.C., Kao,L.G., Chao,C.H., Linacre,A. and Lee,J.C.-I. (1999) Identification of victims of the 1998 Taoyuan Airbus crash accident using DNA analysis. *International Journal of Legal Medicine*, **113**, 43–46.

Jordan,M.I., Ghahramani,Z., Jaakkola,T.S. and Saul,L.K. (1999) An introduction to variational methods for graphical models. *Learning in Graphical Models*. Kluwer Academic Publisher, pp. 105–161.

Lauritzen,S.L. and Spiegelhalter,D.J. (1988) Local computations with probabilities on graphical structures and their application to expert systems. *Journal of Royal Statistical Society (Series B)*, **50**, 157–224.

Lauritzen,S.L. and Sheehan,N.A. (2003) Graphical models for genetic analyses. *Statistical Science*, **18**, 489–514.

Makhorin,A (2001) *GNU Linear Programming Kit*. Moscow Aviation Institute, Moscow, Russia.

Mortera,J., Dawid,J. and Lauritzen,S.L. (2003) Probabilistic expert systems for DNA mixture profiling. *Theoretical Population Biology*, **63**, 191–205.

Olaisen,B., Stenersen,M. and Mevag,B. (1997) Identification by DNA analysis of the victims of the August 1996 Spitsbergen civil aircraft disaster. *Natural Genetics*, **15**, 402–405.

Pearl,J. (1986) Fusion, propagation and structuring in belief networks. *Artificial Intelligence*, **29**, 241–288.

Pearl,J. (1988) *Probabilistic Reasoning in Intelligent System: Networks of Plausible Inference*. Morgan Kaufmann, San Mateo, CA.

Valdes,A.M., Slatkin,M. and Freimert,N.B. (1993) Allele Frequencies at Microsatellite Loci: The Stepwise Mutation Model Revisited. *Genetics*, **133**, 737–749.

Whitaker,J.P., Clayton,T.M., Urquhart,A.J., Millican,E.S., Downes,T.J., Kimpton,C.P. and Gill,P. (1995) Short tandem repeat typing of bodies from a mass disaster: high success rate and characteristic amplification patterns in highly degraded samples. *BioTechniques*, **18**, 402–405.

Xing,E.P., Jordan,M.I. and Russell,S. (2003) A generalized mean field algorithm for variational inference in exponential families. In *Proceedings of the 19th Annual Conference on Uncertainty in AI*.