Available online at www.sciencedirect.com
ScienceDirect
Forensic Science International: Genetics Supplement Series 1 (2008) 664–666
www.elsevier.com/locate/FSIGSS

Research article

Amplification of DNA mixtures—Missing data approach

Torben Tvedebrink [a,*], Poul Svante Eriksen [a], Helle Smidt Mogensen [b], Niels Morling [b]

[a] Department of Mathematical Sciences, Aalborg University, Denmark
[b] Section of Forensic Genetics, Department of Forensic Medicine, Faculty of Health Sciences, University of Copenhagen, Denmark

Received 13 August 2007; accepted 29 August 2007

Abstract

This paper presents a model for the interpretation of results of STR typing of DNA mixtures based on a multivariate normal distribution of peak areas. From previous analyses of controlled experiments with mixed DNA samples, we exploit the linear relationship between peak heights and peak areas, and the linear relations of the means and variances of the measurements. Furthermore, the contribution from one individual's allele to the mean area of this allele is assumed proportional to the average of height measurements on alleles where the individual is the only contributor.

For shared alleles in mixed DNA samples, it is only possible to observe the cumulative peak heights and areas. Complying with this latent structure, we use the EM-algorithm to impute the missing variables based on a compound symmetry model. That is the measurements are subject to intra- and inter-loci correlations not depending on the actual alleles of the DNA profiles. Due to factorization of the likelihood, properties of the normal distribution and use of auxiliary variables, an ordinary implementation of the EM-algorithm solves the missing data problem.

We estimate the parameters in the model based on a training data set. In order to assess the weight of evidence provided by the model, we use the model with the estimated parameters on STR data from real crime cases with DNA mixtures.

Keywords: STR DNA; Mixture; Missing data; EM-algorithm; Compound symmetry model; Multivariate normal distribution

1. Assumptions and definition of the model

We assume to have a two-person STR DNA mixture where we discriminate on the set of loci $\mathcal{S}$ with $|\mathcal{S}| = S$. The amount of DNA contributed to the mixture is modelled by $H^{(k)}$, $k = 1, 2$, which is a weighted mean of observed peak heights with person k as the only contributor where the weights are two for homozygote alleles and one otherwise.

In a mixed DNA trace, it is only possible to observe the vector of cumulative peak areas $\boldsymbol{M}$. The unobservable $4S$-vector of single allelic peak areas $\boldsymbol{A}$ are, however, of interest in order to separate the contributing DNA profiles. We assume independence of the components of $\boldsymbol{A}$ and that they follow a normal distribution. The model assumes proportionality of the mean and variance of $\boldsymbol{A}$, which is expressed as

$$A_{i,s}^{(k)} \sim \mathcal{N}(\alpha_s H^{(k)}, \sigma_s^2 H^{(k)}), \quad i = 1, 2,$$

where α_s and σ_s, $s \in \mathcal{S}$, are loci-dependent parameters. Note that both peak areas (subscript i) for the same person follow the same distribution. Also, the parameterization with $H^{(k)}$ ensures proportionality to the amount of DNA contributed by person k.

The relation on $\boldsymbol{M}$ and $\boldsymbol{A}$ is expressed as a linear transformation T and an additional error term related to the measurement error,

$$\boldsymbol{M} = T\boldsymbol{A} + \boldsymbol{\varepsilon},$$

with $\operatorname{cov}(\boldsymbol{\varepsilon}) = \Sigma$. Let the dimension of $\boldsymbol{M}$ be $n = \sum_{s \in \mathcal{S}} n_s$, where $1 \le n_s \le 4$ is the number of observations on locus s. The transformation T is a $n \times 4S$-block diagonal matrix with diagonal block matrices T_s with 0 and 1 entries according to the profiles in the mixture. For each locus s, we sort the unobservable peak areas $\boldsymbol{A}_s$ by allelic number within each person. If we for locus s assume the mixture is formed by $P_s^{(1)} = (10, 12)$ and $P_s^{(2)} = (9, 12)$, the associated block matrix T_s is given as

$$\boldsymbol{M}_s = \begin{pmatrix} M_{s,9} \\ M_{s,10} \\ M_{s,12} \end{pmatrix} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix} \begin{pmatrix} A_{s,10}^{(1)} \\ A_{s,12}^{(1)} \\ A_{s,9}^{(2)} \\ A_{s,12}^{(2)} \end{pmatrix} + \boldsymbol{\varepsilon}_s,$$

which adds together the entries in $\boldsymbol{A}$ relating to the same allele.

* Corresponding author at: Fredrik Bajers Vej 7G, DK-9220 Aalborg East, Denmark. Tel.: +45 99409981; fax: +45 98158129.
E-mail address: tvede@math.aau.dk (T. Tvedebrink).

1875-1768/$ – see front matter
doi:10.1016/j.fsigss.2007.08.002

In order to cope with the variable amplification abilities across alleles within each locus, we standardize the residual $\boldsymbol{\varepsilon}$ by the observed peak heights $\boldsymbol{h}$ by defining the scaled residual $\tilde{\boldsymbol{\varepsilon}} = (\tilde{\varepsilon}_j)_{j=1}^n = (\varepsilon_j/\sqrt{h_j})_{j=1}^n$, i.e. the variance is proportional to the peak heights. As the number of observations from case to case are likely to vary, we assume a compound symmetry model for the covariance of $\tilde{\boldsymbol{\varepsilon}}$, $\mathrm{cov}(\tilde{\boldsymbol{\varepsilon}}) = \tilde{\Sigma}$, since this structure can be made independent of the present alleles in the mixture.

We parameterize the covariance $\tilde{\Sigma}$ as an additive structure using $\Lambda = \{\nu_{st}\}_{s,t\in\mathcal{S}}$ and $\boldsymbol{\tau} = (\tau_s)_{s\in\mathcal{S}}$ such that ν_{st} is the covariance between two allelic measurements on system s and t, and the variance of a measurement on system s is then $\nu_{ss} + \tau_s$.

2. Implementation of the EM-algorithm

In order to handle the latent structure of $\boldsymbol{A}$ and the associated missing data problem, we use the EM-algorithm to impute the missing observations. However, since the dimensions of $\boldsymbol{M}$ and subvectors hereof vary from case to case, we actually obtain a likelihood which is not very well suited for an implementation of the EM-algorithm.

The problem is solved by introducing appropriate auxiliary variables. This allows for an implementation in the usual full exponential family framework of the EM-algorithm with the constraint of the ν_{ss}-parameters to be positive, i.e. this method imply positive within loci covariances.

The estimators for the parameters included in the model defined in Section 1 can be derived using appropriate orthogonal projections and standard results of normal linear models.

3. Results

The EM-algorithm was implemented in the statistical software R and executed with several different sets of initial values. For each set of initial values, we made 30,000 iterations where the parameters were estimated using a training set consisting of mixtures from controlled experiments conducted at Section of Forensic Genetics, University of Copenhagen[2]. In this data set, only pairwise mixtures of four different profiles were included. Hence, the estimated loci parameters may be biased with respect to the included alleles.

A statistical test based on a χ^2-approximation indicated that $\boldsymbol{\tau} = 0$ with a p-value of 0.9999. Table 1 shows the parameter estimates from the reduced model with $\boldsymbol{\tau} = 0$.

In the Λ part of Table 1, the shading shows the correlations while the remaining partition shows the covariances. We see that most of the loci are highly correlated. This indicates that evaluation of DNA evidence with the assumption of independence across loci is an extensive simplification. Furthermore do the estimates of $\boldsymbol{\alpha} = (\alpha_s)_{s\in\mathcal{S}}$ and $\boldsymbol{\sigma}^2 = (\sigma_s)_{s\in\mathcal{S}}$ support the assumption of proportionality of mean and variance of $\boldsymbol{A}$. Further analysis shows the factor of proportionality is close to 190.

The different amplification behaviour of dye band from the fluorescent reaction is also identifiable in the parameter estimates. The larger amplification of the green dye band and smaller amplification of the yellow dye band is captured in the parameter estimates of α_s. In Table 1, we see the ordering of the α s related to the yellow band is less than the blue band, which again is smaller than the green band.

In order to monitor the convergence of the EM-algorithm, we computed the deviance after each iteration. After 1100 iterations, the improvement for successive deviances were less than 0.01. Models with Λ initiated as a zero-matrix showed the worst fit of all sets of initial values.

In addition to the parameter estimates and deviance, we also computed the asymptotic variances of the estimates by the normality approximation of the MLE with the inverse Fisher information as covariance matrix. We found the estimated standard deviation of both $\boldsymbol{\alpha}$ and $\boldsymbol{\sigma}^2$ indicated reasonable good

Table 1
Parameter estimates after 30,000 iterations of EM-algorithm

	Yellow dye band			Blue dye band				Green dye band		
	FGA	THO	*D19*	*D2*	*v*WA	*D3*	*D16*	*D21*	*D8*	*D18*
FGA	1151.536	773.909	1441.000	1492.006	1044.479	857.456	1305.358	1033.495	397.344	1461.591
THO	0.806	925.693	1042.031	1090.164	587.580	664.553	527.211	654.255	582.881	1085.372
D19	0.866	0.925	2052.831	2151.759	1319.491	1050.949	1716.853	1279.925	619.224	1964.678
D2	0.848	0.890	0.953	2481.701	1438.362	1237.071	1848.140	1354.935	765.346	2077.812
Λ = *v*WA	0.890	0.936	0.915	0.881	1082.097	821.782	1339.643	975.914	340.353	1449.906
D3	0.742	0.829	0.789	0.845	0.856	864.066	954.082	776.642	536.557	1142.766
D16	0.197	0.558	0.477	0.536	0.431	0.637	1915.776	1201.832	246.497	1653.911
D21	0.879	0.919	0.937	0.883	0.987	0.860	0.409	952.266	380.722	1353.252
D8	0.396	0.761	0.756	0.719	0.697	0.743	0.669	0.750	820.995	750.020
D18	0.930	0.940	0.885	0.878	0.961	0.850	0.361	0.936	0.587	2196.651
α =	5.529	5.992	6.149	7.014	7.642	8.248	9.102	8.918	10.189	10.175
σ^2 =	596.532	730.285	1,002.926	1,236.310	1,146.784	1,331.789	1,821.040	1,797.387	1,854.943	3,208.726

The Λ-matrix show the covariances and correlations (shaded).

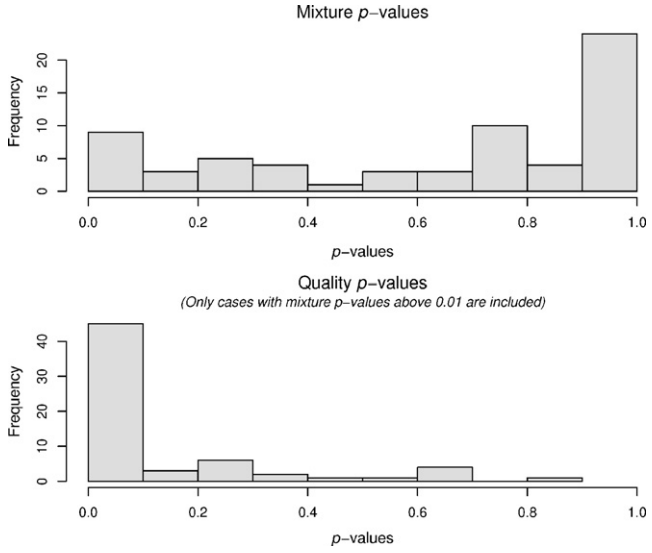

Fig. 1. Histograms of Mahalanobis distances.

estimates of these parameters. Large asymptotic standard deviations of Λ did, however, indicate the possibility of model reductions.

4. Discussion

For evaluating data from real crime cases, we use the estimated parameters from Table 1. The Mahalanobis distance of $\boldsymbol{M}$ with respect to the model evaluates the evidence in one step. This give rise to problems to real crime case data since degradation of DNA and other unbalances are not incorporated in the model. However, by conditioning on the loci-sums $\bar{\boldsymbol{M}}$, we can assess the goodness of fit for each system. Using the estimated parameters from the EM-algorithm, the Mahalanobis distances have approximate χ^2-distributions:

$$(\boldsymbol{M} - \boldsymbol{\mu}_{\boldsymbol{M}|\bar{\boldsymbol{M}}})^\top \operatorname{cov}(\boldsymbol{M}|\bar{\boldsymbol{M}})^- (\boldsymbol{M} - \boldsymbol{\mu}_{\boldsymbol{M}|\bar{\boldsymbol{M}}}) \sim \chi^2_{n-S}$$
$$(\bar{\boldsymbol{M}} - \boldsymbol{\mu}_{\bar{\boldsymbol{M}}})^\top \operatorname{cov}(\bar{\boldsymbol{M}})^{-1} (\bar{\boldsymbol{M}} - \boldsymbol{\mu}_{\bar{\boldsymbol{M}}}) \sim \chi^2_S.$$

If the former test shows a reasonable fit to the model assumptions, e.g. a p-value above 0.01, the latter model for the loci-sums is valid. Hence, the latter test gives indications of the sample quality, e.g. low p-values may indicate the DNA material is degraded.

In Fig. 1, we have plotted histograms of the two p-values for 66 real crime cases made available by Section of Forensic Genetics, University of Copenhagen[2]. The lower part of Fig. 1 shows that 35 cases have a p-value for the quality of the sample less than 0.01. Furthermore does the top panel of Fig. 1 indicates that the model may be to simple and needs further investigation.

5. Further details

A more detailed description of the model, the implementation of the EM-algorithm with full source code, data plots and further discussion of the results are available at http://www.math.aau.dk/~tvede.

Conflict of interest

None.