Validation of an STR peak area model

Robert G. Cowell *

Faculty of Actuarial Science and Insurance, Sir John Cass Business School, City University London, 106 Bühnill Row, London EC1Y 8TZ, UK

ABSTRACT

In analyzing a DNA mixture sample, the measured peak areas of alleles of STR markers amplified using the polymerase chain-reaction (PCR) technique provide valuable information concerning the relative amounts of DNA originating from each contributor to the mixture. This information can be exploited for the purpose of trying to predict the genetic profiles of those contributors whose genetic profiles are not known. The task is non-trivial, in part due to the need to take into account the stochastic nature of peak area values. Various methods have been proposed suggesting ways in which this may be done. One recent suggestion is a probabilistic expert system model that uses gamma distributions to model the size and stochastic variation in peak area values. In this paper we carry out a statistical analysis of the gamma distribution assumption, testing the assumption against synthetic peak area values computer generated using an independent model that simulates the PCR amplification process. Our analysis shows the gamma assumption works very well when allelic dropout is not present, but performs less and less well as dropout becomes more and more of an issue, such as occurs, for example, in Low Copy Template amplifications.

© 2009 Elsevier Ireland Ltd. All rights reserved.

1. Introduction

Analysing mixed DNA STR profile samples is acknowledged to be a complex task [3]. The importance of using peak-area information (or, alternatively, peak height information: peak area values are known to be highly correlated with peak height values) to analyse mixed DNA STR profile samples is recognised. In recent years, a variety of methods has been proposed to exploit peak area information, see for example [4–6]. This paper examines the probabilistic model developed and applied by Cowell, Lauritzen and Mortera in a series of recent papers [1,7,8]. In particular, we focus on checking the assumption of using Gamma distributions to model the probabilistic variation in peak area values obtained from the PCR process. Ideally, this would be carried out using data obtained from an extensive controlled experiment involving amplifications of many DNA samples. However, this is not practical. Instead we opt for a simulation approach, in which we simulate peak area values using the stochastic model in [2].

The plan of the paper is as follows. The next section summarises the mathematical model presented in [1] and the simulation model of [2], but ignoring the complication of stutter. Then we examine the fit of the gamma model to simulated data, and also make some analytic comparisons. We then summarise our results. A set of Appendices containing technical mathematical proofs is available as an online supplement.

2. Models and methodology

In this section, we present a brief summary of the gamma model of [1], a summary of the PCR simulation model of [2], and a discussion of the application of the simulation model to testing the gamma model.

2.1. Summary of peak area model for mixtures

The gamma model of [1] considers $I$ potential contributors to a DNA mixture. Let there be $M$ markers to be used in the analysis of the mixture, and that marker $m$ has $A_m$ allelic types, $m = 1, \ldots, M$. Let $\theta_i$ denote the proportion of DNA from individual $i$ prior to PCR amplification, with $\theta = (\theta_1, \theta_2, \ldots, \theta_I)$ denoting the vector of proportions from all contributors. It is assumed that this pre-amplification proportion of DNA is constant across markers.

For a specific marker $m$, the peak weight $W_{i,a}$ is the peak area at allele $a$ multiplied by the allele number. This is a simple correction for preferential amplification of alleles. The model is idealized in that it ignores complicating artifacts such as stutter,
drop-out alleles and so on, and makes the following further assumptions:

- If \( W_{ia} \) denotes the contribution of individual \( i \) to peak weight at allele \( a \), then
  \[
  W_{i,a} = \sum_{i} W_{ia}.
  \]
- \( W_{i,a} \) is approximately proportional to the amount of DNA of type \( a \).

The key modelling assumption is that each contribution \( W_{ia} \) from individual \( i \) to peak weight at allele \( a \) is assumed to have a Gamma distribution:

\[
W_{ia} \sim \Gamma(\rho \gamma_i n_{ia}, \eta)
\]

where

- \( \gamma_i \) is the amount of DNA from individual \( i \) in mixture;
- \( n_{ia} \) is the number of alleles of type \( a \) carried by individual \( i \);
- \( \eta \) determines scale and \( \rho \) is the amplification factor. Both may be marker dependent.

It follows from the gamma distribution assumption that

\[
W_{i,a} = \sum_{i} W_{ia} \sim \Gamma(\rho \sum \gamma_i n_{ia}, \eta)
\]

and

\[
W_{i,a} = \sum_{a} W_{i,a} = \sum_{a} \sum_{i} W_{ia} \sim \Gamma(2\rho \gamma, \eta).
\]

By scaling the weight of each allele by the total for the marker we obtain relative weights \( R_a = W_{i,a}/W_{i,a} \), where it follows that

\[
R_a = \text{Dir}(\rho B_a),
\]

where \( B_a = \sum \gamma_i n_{ia} \) is the weighted allele number, and \( B_a = 2\gamma \) is twice the total amount of DNA \( \gamma \) and is marker independent. Note that

\[
E(R_a) = \mu_a = \frac{B_a}{\rho} = \sum \frac{\theta_i n_{ia}}{2},
\]

\[
\text{Var}(R_a) = \mu_a(1 - \mu_a) = \sigma^2 \mu_a(1 - \mu_a)
\]

where we define \( \sigma^2 = 1/(2\rho \gamma + 1) \). Using these moments, a moment-matching approximation to the gamma model using normal distributions can also be developed; this is described in [7].

The attractiveness of the gamma model, and its normal approximation, is that it provides a framework in which a variety of complex mixture problems can be both formulated and solved using Bayesian network technology.

### 2.2. Summary of simulation model (excluding stutter)

Building on the binomial model of [9], Gill et al. [2] developed a simulation model of the process of analysing STRs. In this section we present a simplified version of their model that excludes stutter. (In Appendix C we present their extended model with a correction for dealing with stutter, and also including dropout occurring through silent alleles.)

The simulation model of PCR amplification without stutter is much simpler than the one with stutter, because it is possible to treat each allele independently of other alleles. Thus, let us concentrate on a sample containing \( N \) alleles of type \( a \) bound within cells. The simulation model then has the following three steps that are readily implemented in a small computer program:

#### Step 1: Extract the alleles from the cells:

It is assumed that each allele \( a \) has a probability \( \pi_{\text{extraction}} \) of surviving this process, independently of other alleles, hence the total number of type \( a \) alleles \( n \) has a binomial distribution:

\[
n_0(n) \sim \text{Binom}(n, \pi_{\text{extraction}}).
\]

#### Step 2: A portion \( \pi_{\text{aliquot}} \) of the extracted sample is used for PCR amplification. It is assumed that an individual allele that has survived the extraction process will be randomly selected with probability \( \pi_{\text{aliquot}} \) for amplification, independently of other alleles, hence the total number of available for amplification, \( n_0 \), has a binomial distribution:

\[
n_0(n) \sim \text{Binom}(n, \pi_{\text{aliquot}}).
\]

Note that this may be combined with the distribution in Step 1, to give

\[
n_0(n) \sim \text{Binom}(n, \pi_{\text{extraction}} \pi_{\text{aliquot}}).
\]

#### Step 3: The surviving alleles are then subject to \( K \) cycles of amplification.

During each amplification cycle, each allele is either duplicated with probability \( \pi_{\text{PCR eff}} \) or not duplicated with probability \( 1 - \pi_{\text{PCR eff}} \). Assuming that alleles do not interfere with each other in the amplification process the number of alleles after the \( k \)th cycle, and that the replication probability \( \pi_{\text{PCR eff}} \) is the same for each cycle, then \( n_k \) is related to the number \( n_{k-1} \) in the previous cycle by the following recurrence relation:

\[
n_k = n_{k-1} + \text{Binom}(n_{k-1}, \pi_{\text{PCR eff}}).
\]

For standard amplifications, \( K \) is set to 28 cycles, whilst for low copy number procedures, appropriate with \( N < 20 \), \( K = 34 \) simulated cycles are used. Using experimental data based on their laboratory procedures, [2] estimated the following values for the various probability parameters \( \pi_{\text{extraction}} = 0.6 \), \( \pi_{\text{aliquot}} = 20/66 \) and \( \pi_{\text{PCR eff}} = 0.82 \). We use these values in all our simulations. Finally, the alleles register a peak if their number exceeds the threshold total of \( 2.35 \times 10^7 \), and numbers are scaled by the factor of \( 2 \times 10^7 \) to convert them to a representative peak area value.

Note that the iteration in Step 3 defines a Galton-Watson branching process. Such processes have been used to model aspects of PCR by other authors, for example [10] who studied mutations in the PCR process, and [11] who estimate the efficiency of the amplification process.

### 2.3. Nature of simulations

There are some differences in the simulation model and the gamma model that need to be accommodated in order to use the simulated peak areas to test the gamma model. The gamma model introduces weighted areas, that is, the weight associated with a peak area of an allele having repeat number \( r \) is the peak area times \( \sigma^2 \mu_a(1 - \mu_a) \). Using these moments, a moment-matching approximation to the gamma model using normal distributions can also be developed; this is described in [7].
height has to be above a certain threshold to be counted as a peak. Hence in simulating the peaks areas we shall set the threshold number to zero, and merely scale the counts by $2 \times 10^6$ to obtain an area value.

This calibration factor, converting simulated counts to area values, is not ideal because it does not give the correct range of values for peak areas that one would expect to see in terms of RFU units from an electropherogram. However, our analysis will not depend on the precise value that the calibration should take for two reasons. The first is that a positive multiple of a gamma distributed random variable is also a gamma distributed random variable. Hence in testing whether the gamma distribution family is suitable to use to model peak areas, the value of the calibration factor is not an issue. The second reason why the value of the calibration factor is not important for this paper is that the factor is not an issue. The second reason why the value of the calibration factor is not an issue is that the value of the calibration factor is not an issue.

In the simulation model, dropout of allele $a$ occurs either because no alleles of type $a$ survive the first two stages of the simulation – extraction and subsequent sampling – or if some do they fail to amplify beyond the threshold. As we are setting the threshold to zero, this means that dropout can only happen in our simulations if an allele of a certain type $a$ fails to survive the first two steps of the simulation. Therefore, in simulating many samples, we shall only retain those in which dropout does not occur. We shall return to these issues in more detail in Section 4.

It is worthwhile emphasising here the distinct natures of the gamma model and the simulation model. The simulation model of [2] generates mixture peak areas that appear to capture the important statistical characteristics of peak area values obtained from PCR amplification of real DNA mixture samples, incorporating artifacts such as stutter and dropout. The simulation method itself is not used to analyse any given mixture, for example to identify unknown genetic profiles of individual contributors. In contrast, the gamma model has been developed for the purpose of analysing given mixtures using the peak area information. Calculations based on the gamma model can be carried out quickly and accurately using probabilistic expert system software. The gamma model presented and analysed in this paper does not cater for stutter or dropout, so where such effects start to become dominant in the simulated samples we could expect the gamma model to start breaking down, and indeed we shall see that this is the case. Work is in progress to extend the gamma model to cater for artifacts, and it is hoped to present this extension elsewhere. The aim of this paper is to see if the gamma model provides a good fit to situations in which artifacts are not present: if it does then it makes sense to try and extend the model to cope with artifacts. On the other hand if the gamma model turns out to fit poorly in the simplified situation of no artifacts, then one would not expect an extended gamma model to fit any better. As we shall see, the gamma model fits well when artifacts are not present, justifying the attempt to extend it to cater for artifacts.

3. Results

Given the restrictions described above, we proceed to use the simulation model to test the following aspects of the gamma model:

peak areas: The peak areas follow gamma distributions with mean proportional to the amount of DNA in the sample, and scale parameter independent of the amount of DNA in the sample.

Relative areas: The gamma model predicts that the relative areas $R_a = W_{aa}/W_{aa}$ follow Dirichlet distributions $R_a \sim \text{Dir}(\beta R_a)$, where $R_a = \gamma \sum_i h_i n_{ai}$. For the special case of amplifying the DNA of a single heterozygote person, the prediction simplifies to $R_a \sim \text{Beta}(2,\gamma)$ having mean $E(R_a) = 0.5$ (because no allowance has been made for preferential amplification) and variance $V(R_a) = 1/(4\gamma(\gamma + 1))$. We examine how well the Beta distribution fits, and also examine the behaviour of the variance: according to the model the inverse of the variance will be linearly proportional to the amount $\gamma$ of DNA prior to amplification.

We shall look at each aspect in turn.

3.1. Peak area size

In Fig. 1 is plotted the results of simulating counts (representing areas) for three different starting values of $N$, the number of alleles...
so a certain type. For each value of $N$, 10,000 amplifications were simulated, and those that did not lead to dropout were retained. The histogram plots in the first column indicate more skewness in the distribution the lower the value of $N$. All indicate that gamma distributions are appropriate. This is confirmed in the second column of plots, in which moment estimation was used to estimate the parameters of the gamma distribution for each $N$, and values simulated from this to construct the quantile–quantile plots in the second column. All plots are highly linear over the range except for departures from linearity apparent to the top right hand corners of the plots. In [7] a normal approximation to the gamma model was suggested. Hence in the third column is plotted normal quantile–quantile plots based on the simulated counts. All plots are linear over the range (−2, 2), but departures are apparent outside this range for the two simulations with starting values $N = 5$ and $N = 20$, consistent with the skewed histograms. It is worth pointing out that the plots in Fig. 1 were based on 28 simulated PCR cycles, but starting values 34 simulated cycles should be used [2]. However, using 34 simulated cycles leads to similar plots (not shown) to those in Fig. 1.

A second assumption of the gamma model is that the mean area associated with an allele is proportional to the amount of DNA of that allelic type in the sample prior to amplification: $W_a = \Gamma(\gamma/n_a, \eta)$. This assumption is verified by the plot in Fig. 2 (10,000 simulations for each integer value of $N \in (1, 100)$) which shows a strong linear dependence of mean area on the amount over a large range, with just a hint of departure from linearity in the low copy number regime. This result can be derived analytically for the simulation model, as shown in Appendix A.

Another assumption of the gamma model is that, for each marker, the scale parameter $\eta$ does not depend on the amount. To test this, scale parameters were estimated for each value of $N$ using the same simulations as for Fig. 2. The plot in Fig. 3 shows that the common scale parameter assumption appears reasonably valid for the range $N > 25$, but the assumption appears to breakdown in the range of values of $N$ in the low copy number regime. The breakdown is arising in the variation arising from the selection of alleles in the pre-PCR stage. For the PCR stage itself the assumption of constant scale parameter $\eta$ holds as follows. It is shown in Appendix A that the mean number of alleles arising from $r$ simulated PCR cycles starting from a single allele is $E_r[X] = (1 + \pi_{PCRl})^r$, and the variance is $V_r[X] = (1 - \pi_{PCRl})(1 + \pi_{PCRl})^{-1}(1 + \pi_{PCRl})^{-2} - 1$. For an initial

$$\frac{V_r[X]}{E_r[X]} = \left(\frac{1}{N} + \frac{1}{N^2}\right) - 1$$

number $N$ alleles amplifying independently, both the mean and variance are multiplied by $N$. (This follows because of the independence. Alternatively, it can be found by finding the moments using $F_r(t)^N$.) For the gamma distribution, the moment matching estimate of the scale parameter $\eta$ is $E_r[X]/V_r[X] = (1 - \pi_{PCRl})(1 + \pi_{PCRl})^{-1}/(1 + \pi_{PCRl})$ which does not depend on $N$.

### 3.2. Relative area size

We have seen that the use of gamma distributions appears to be justified from the simulations in Section 3.1. We now look at the gamma model prediction that the relative peak areas obtained by amplifying the alleles of a heterozygote individual are Beta distributed. In Fig. 4 we plot the results of simulations of relative areas for three different starting values for the number of alleles of each type. Each simulation sampled 10,000 amplification values, only those that did not result in a dropout were retained for calculating relative area values. To the left are histograms, indicating a symmetric distribution with mean 0.5. On the right are corresponding quantile–quantile plots which use the simulated values for actual quantiles, plotted against quantiles of values simulated from a beta distribution of mean 0.5 and variance matching that of the relative area values. All quantile–quantile plots yield close to straight lines, indicating that the beta distribution is a very good fit to the data, even for the low-copy number regime with a starting value of 5 for each type of allele.

Fig. 5 shows the results of a similar simulation, in which the starting number of alleles was unequal, in a ratio $n_a : n_b = 1 : 3$. This could correspond to two homo-zygote contributors contributing DNA in a ratio of 1:3 to a mixture, or one homo-zygote and one heterozygote having one allele in common and contributing equal amounts to a DNA mixture. Again, the quantile–quantile plots fit straight lines well, indicating that the relative areas are following appropriate Beta distributions.

A second prediction of the gamma model regarding the relative areas is that the inverse of the variance of the relative area is linearly proportional to the starting amount of DNA. From the histogram plots in Fig. 4 the variance does appear to decrease as $n_a$ increases. For a fuller analysis of the dependence of variance on amount, 100 amplifications were simulated of a heterozygote, with each of the starting values $n_a (= n_b) = 1, 2, \ldots, 100$ for the number of alleles to each type, with 28 simulated PCR cycles. From each simulation the variance of the relative areas was calculated for
those simulated amplifications that did not lead to dropout. The variance and inverse variance are plotted against $n_a$ (which is proportional to the amount of DNA in the gamma model) in Fig. 6.

From the plot of inverse variance, a good straight line fit is apparent for the range $n_a \geq 20$ (the sample correlation coefficient of the values in this range is 0.998). Thus, the gamma model is providing an excellent fit over this range. However, it is clear than for smaller amounts the gamma model prediction regarding the dependence of variance on amount is breaking down. The plot on the right reaches a minimum for $n_a = 14$, and for lower values of $n_a$ the inverse variance is increasing, corresponding to the variance decreasing with decreasing $n_a$ for $n_a \leq 14$.

This somewhat surprising inflexion in the curve can be understood by a closer examination of how variability arises from the simulation model. Recall there are two stages to the simulation process, the pre-PCR selection stage (consisting of two steps) and

---

Fig. 4. Simulations of relative areas for amplification of hetero-zygotic individual, for three different starting values of the number of alleles of each type.

Fig. 5. Simulations of relative areas for amplification for a 1:3 imbalance of number of alleles of either type to start with.
the PCR stage. Each stage provides variability. In the pre-PCR selection stage, alleles are sampled for the PCR stage, each with probability \( p = 0.6/3.3 \approx 0.182 \). Thus an equal number \( n_a = n_b \) of alleles of types \( a \) and \( b \) can easily become unbalanced prior to the PCR stage of the simulation, and the PCR amplification stage can further exacerbate the in-balance.

Now consider the extreme case where we start out with \( n_b = n_a = 1 \). The simulations only retain those values where there is no dropout. This requires that both alleles are present for the simulation of the PCR stage, and thus start out with the same number, that is \( n_a = n_b = 1 \), and the ratio of one allele count to the total is 0.5. Thus there is no pre-PCR variability in the relative amounts of the two alleles entering the amplification stage, and so the total variability is less than may be expected at first sight. On the other hand, for very large values of \( n_a = n_b \), the Beta distribution become very highly peaked and narrow, so that the values sampled from \( \text{Binom}(n_a, p) \) and \( \text{Binom}(n_b, p) \) are close to \( n_a p \) and \( n_b p \), leading to a relative values close to 0.5 and again a small variance that goes to zero as \( n_a = n_b \to \infty \). For the intermediate values of \( n_a = n_b \), both sources of variability are present, and so by continuity we should expect a maximum value somewhere. For the parameters used in the simulation the greatest variance occurs at \( n_b = 14 \), as shown in Appendix B for the pre-PCR sampling stage, and also \( n_b = 14 \) in Fig. 6 where both sources of variation are combined.

4. Discussion

This paper has presented the results of a simulation study of the gamma model of [1] for the distribution of peak areas arising in the PCR amplification of STRs. The study has been based on a simplified version of the model of [2] for simulating the whole process of amplifying STRs to create realistic simulations of peak area distributions. On the whole, the gamma model is seen to compare favourably for modelling the peak areas simulated from the model of [2]. The plots in Fig. 1 show that gamma distributions model the simulated areas well, even in the low copy number regime \((N < 20)\). The normal approximation suggested in [7] works well for large \( N \) but breaks down in the low copy number regime. We have shown in Appendix A that the assumption of the linear dependence of the mean of the gamma distribution on the amount is a consequence of the simulation model. There is a slight departure for the low copy number regime that can be explained by the pre-PCR sampling of alleles in the extraction and aliquot selection processes that can lead to some or all alleles of specific type not being selected for amplification. The gamma model does not cater for such dropout behaviour. The assumption of the common scale parameter \( \eta \) has been shown to hold via simulation for large \( N \) values but appears to break down in the low copy number regime. Again this can be traced to the pre-PCR sampling of alleles for the amplification process. The results of Section 3.2 shows that the relative areas of two alleles in a simulated amplifications closely follow beta distributions even in the low copy number regime. The prediction of the gamma model that the variance of the relative areas should increase as the amount of DNA decreases holds in the regime for large \( N \), but breaks down in the low copy number regime. The explanation of this behaviour in the simulation model is given in detail in Appendix B, and again lies in the variability of alleles introduced by the pre-PCR sampling of alleles for amplification.

We must now return to the points made in Section 2.3, and address the relevance of the simulation study of this paper to the application of the gamma model to the analysis of real STR mixtures.

The main issue is that the gamma model of [1] used a simple correction for preferential amplification that the present simulation study did not use, because the simulation model of [2] does not include simulation of preferential amplification. The simple correction of [1] does not take into account the dependence of preferential amplification of DNA on the condition of the sample. The question arises as to whether it provides a good enough approximation: ultimately this can only be resolved by experimental analysis of real STR PCR amplifications.

Another issue concerns stutter. The simulations reported in this paper used a simplified model that ignored the generation of stutter peaks. This was because the model of [1] does not model stuttering. However, an extension to take account of stuttering is being developed and will be reported elsewhere. Another property of the simulations used this paper is that they were conditional on no dropout occurring. This was appropriate for testing the model of [1] as it does not take account of dropout. Real PCR amplifications do lead to dropout, and there appear to be three main causes. These are as follows:

1. Pre-selection of alleles lead to some alleles types not being selected for the PCR amplification stage. This is particularly acute for low copy number amplification.
2. Silent alleles: alleles that do not amplify because the primer does not bind to the flanking regions of the DNA.
3. Alleles are amplified, but below a threshold for detection.
Of these possibilities, extensions of the gamma model to incorporate dropout due to (1) and (2) are being developed and will be reported elsewhere. The problem of modelling the detection threshold has still to be addressed within the context of the (extended) gamma model.

In summary, the gamma model provides a good approximate fit to data generated using the simulation model when the initial number of alleles $N$ is relatively large so that the stochastic variation introduced by the selection steps 1 and 2 described in Section 2.2 are much smaller that arising from the amplification process of step 3. In contrast the fit of the gamma model breaks down in the low copy number regime, and the lack of fit can be traced to the pre-PCR selection stage of the simulation model which is not captured by the gamma model. We agree with Gill et al. [2] that this pre-PCR selection stage can be a dominant source of peak imbalance and dropout in real DNA samples when the initial number of alleles in a DNA sample is initially small. However, extensions of the gamma model to incorporate dropout to take account of this are being developed. The present study has shown that the simple gamma model of [1] analysed in this paper forms a solid foundation upon which to build such extensions.

Acknowledgements

The author would like to thank Professors S. L. Lauritzen and J. Mortera for helpful discussions and encouragement regarding this work, and two referees whose suggestions have improved the paper.

Appendix A. Supplementary Data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.fsigen.2009.01.006.

References