A BAYESIAN APPROACH TO ASSESSING LAB PROFICENCY WITH QUALITATIVE PCR ASSAYS USED TO DETECT BIOTECH TRAITS IN CROP SEED

Kirk M. Remund

Glenn D. Austin

A BAYESIAN APPROACH TO ASSESSING LAB PROFICENCY WITH QUALITATIVE PCR ASSAYS USED TO DETECT BIOTECH TRAITS IN CROP SEED

Kirk M. Remund
Glenn D. Austin
Monsanto Company
800 North Lindbergh Blvd.
St. Louis, MO 63167

Abstract

Many seed testing laboratories currently use polymerase chain reaction (PCR) assays to test conventional crop seed for the adventitous presence of biotech trait seed. Seed organizations and companies are spending much time and resources assessing laboratories proficency in running PCR assays. Since many of these assays provide qualitative rather than quantitative results, laboratories must go through a significant effort to obtain adequate assay error estimates. Many sample-processing steps are very similar from assay to assay and therefore error results from different assays may be combined using a Bayesian approach to obtain estimates of assay error rates with increased precision. This paper introduces a relatively simple Bayesian approach that can be used to combine data from a present assay of interest with prior lab data on related assays to obtain updated estimates of assay error rates. If this approach is successfully implemented it can yield as much as a ten-fold reduction in required testing resources.

1. Introduction

Many laboratories test seed and plant material for the unintended or intended presence of transgenic events. Many of these laboratories are using qualitative PCR methods (either end-point Taqman® or gel-based PCR). Presently each lab must demonstrate proficency on each assay before the lab can use the assay for production samples. False-positive (FPR) and false negative (FNR) rates are the result of this process. A labs assay error rates must be within specified ranges to pass for use on production samples. This lab proficency testing effort has required a large amount of time and resources. The objective of this proposal is to reduce the effort required for this lab proficiency testing process.

The present lab proficency testing process is briefly described here. There are three major steps to the lab proficiency testing process for qualitative PCR assays. First, a primary approval kit is sent to the lab that is generally comprised of a specified number of positive and negative blind samples for a given assay. To pass the primary approval, a lab must have a very small number of misclassifications of each type. After primary approval, a lab is tentatively cleared to test production samples but must enter a next phase of proficency testing (called Phase 1). In this phase, a lab must test a much larger set of positive and negative blind samples spread over at least three time points. At the completion of Phase 1 testing, a lab is given approval to test production samples if they meet the testing requirements for the specific assay. A final step in

the lab proficency testing process is to take part in recurrent testing program such as a round robin or ring test with other labs.

There have been a number of challenges with the present lab proficency testing process which include 1) obtaining enough adequate reference material, 2) large resource requirements for participating labs and 3) large time and resource commitment by lab preparing sample kits.

Many of the testing laboratories have been involved with much lab proficency testing on several PCR assays over the last two to three years. We have noted that many of these assays have similar error rates. We believe that this is largely due to the similarity in the testing steps from one assay to another (i.e., sample preparation, extraction and PCR steps are similar). We have gained enough process information and data from some labs to have confidence in a lab's overall ability to run PCR assays. The focus of this paper is to use this data and information on a lab's prior proficency testing efforts along with present data from the assay of interest to obtain increased confidence in the assay error estimates at that lab.

2. Comparing Historical and Present Data

For the historical data to be considered with the present assay testing data, a process checklist will be used to ensure that each historical datasets came from virtually the same process steps as those used to generate the data on the present assay being considered. Such steps as sample grinding procedures and primer quality programs are included in this process checklist. If the processes are judged to be sufficiently similar for the present and historical assays, then it is reasonable to combine datasets.

3. Error Rate Estimates and Bounds for Bayesian Approach

Here are definitions for a number of quantities that will be used in this section and the section that follows. They are

x is the number of false-positives or false-negatives from the present data on the assay of interest,

n is the number of blind samples of a given type (i.e., negative or positive) used in the lab proficency test for the present assay,

y_i is the number of false-positives or false-negatives from the ith historical dataset that relates to the present assay under consideration,

m_i is the number of blind samples of a given type (i.e., negative or positive) in the ith historical dataset that relates to the present assay under consideration,

θ is the true unknown FNR or FPR for the present assay under consideration,

α, β are unknown hyperparameters from the beta prior distribution that will be described below and

$\hat{\theta}, \hat{\sigma}^2_\theta$ are mean and variance calculated using the individual error rates for each historical dataset.

Gelman, et al. (1995) as well as Carlin and Louis (1997) are used as a basis for the equations presented below. There are three statistical distributions that are part of this Bayesian approach. The first distribution used is a binomial distribution, which is a function of the random variable x given the parameters n and θ. This distribution is often called the likelihood density and can be expressed as

$$p(x \mid \theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x} \tag{3.1}$$

for all possible x values in the range from 0 to n.

This distribution is used to compute error rate probabilities and confidence limits based solely on the present data available for the assay of interest (e.g., data on the present assay under consideration).

The second statistical distribution used is a beta distribution, which is a function of the parameter θ with hyperparameters α and β (both of these hyperparameters must be greater than zero). This distribution is called the prior density and can be expressed as

$$p(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1} (1-\theta)^{\beta-1} \tag{3.2}$$

where θ is defined on the interval from zero to one (i.e., 0% to 100% error).

The beta distribution in (3.2) is used here due to its conjugate relationship with the binomial distribution. The prior density is used to incorporate the information from the historical data. In many Bayesian applications the hyperparameters α and β in the prior density are specified using subjective information about the historical process. In this application we use the historical data to obtain estimates for these hyperparameters using the following steps. First we obtain the estimates of E(θ) and V(θ) using the calculated mean and variance of the individual historical datasets error rates. Method of moments estimates of α and β can then be obtained by using these estimates in place of E(θ) and V(θ) respectively in the following system of equations

$$\alpha+\beta=\frac{E(\theta)[1-E(\theta)]}{V(\theta)}-1$$
$$\alpha=(\alpha+\beta)E(\theta) \tag{3.3}$$

and solving for α and β. This system of equations and approach is presented in Appendix 3 of Gelman, et al. (1995). The method of moment estimators for α and β are respectively

$$\hat{\alpha}=\frac{\hat{\theta}^2-\hat{\theta}^3-\hat{\theta}\hat{\sigma}_\theta^2}{\hat{\sigma}_\theta^2},$$
$$\hat{\beta}=\frac{\hat{\theta}-2\hat{\theta}^2+\hat{\theta}^3+\hat{\theta}\hat{\sigma}_\theta^2-\hat{\sigma}_\theta^2}{\hat{\sigma}_\theta^2}. \tag{3.4}$$

Alternately, maximum likelihood estimators for these hyperparameters could be considered and might have better properties than the method of moment estimators in (3.4). The estimators in (3.4) are used in this paper because of their simple closed form. Using the likelihood and prior densities, the posterior density becomes

$$p(\theta\mid x)=\frac{\Gamma(\hat{\alpha}+\hat{\beta}+n)}{\Gamma(\hat{\alpha}+x)\Gamma(\hat{\beta}+n-x)}\theta^{\hat{\alpha}+x-1}(1-\theta)^{\hat{\beta}+n-x-1}$$
$$\text{or otherwise stated } \theta\sim\text{Beta}(\hat{\alpha}+x,\hat{\beta}+n-x). \tag{3.5}$$

The posterior density mean is used to estimate the error rate, which uses the historical and present data together. This estimated mean is

$$\hat{\theta}=\frac{\hat{\alpha}+x}{\hat{\alpha}+\hat{\beta}+n}. \tag{3.6}$$

Quantiles from the posterior density can be used to create posterior intervals or upper posterior limits.

4. Obtaining Error Rate Estimates and Bounds Using Present and Historical Data

All historical datasets must pass the process checklist requirements before they can be considered in these Bayesian calculations. The Bayesian approach will provide more appropriate wider error rate bounds than simply pooling data together across datasets when there is much variability in error rate estimates from one dataset to another. As an example, consider a case with the following results:

- present dataset: 40 samples with 2 misclassifications
- historical dataset: 40 samples with 1 misclassification

- historical dataset: 40 samples with 0 misclassifications
- historical dataset: 40 samples with 2 misclassifications
- historical dataset: 90 samples with 9 misclassifications
- historical dataset: 90 samples with 2 misclassifications.

In this example, the historical dataset with 90 samples and 9 misclassifications is quite different than the other datasets. The Bayesian approach yields an estimated error rate of 4.9% and an upper 95% bound of 9.7%. If the datasets error rates are simply pooled then the estimated error rate is 4.7% with an approximate 95% upper bound of 7.1%. This approximate upper bound was taken from Johnson et al. (1993). Since the pooled approach ignores the additional variability that this historical dataset adds, the upper bound are lower.

5. Final Comments

This Bayesian approach may be an effective way to help eliminate the lab proficency testing challenges discussed in the introduction for the testing approach presently used. There is a good possibility that with this approach fully implemented new assays tested by a lab may only require passing an initial kit to receive full approval to run the assay on production samples (i.e., phase 1 can be eliminated).

Acknowledgements

We appreciate the valuable comments from Changjian Jiang and the referee on this paper.

References

Gelman, A., Carlin J.B., Stern, H.S., Rubin, D.B., Bayesian Data Analysis, (1995) Chapman & Hall, New York.

Carlin, B.P., Louis, T.A., Bayes and Empirical Bayes Methods for Data Analysis, (1997) Chapman & Hall, New York.

Johnson, N.L., Kotz, S., Kemp, A.W., Univariate Discrete Distributions, (1993) John Wiley & Sons, New York.

Lindley, D.V., Introduction to Probability and Statistics from a Bayesian Viewpoint, Part 2. Inference, (1st Edition) Cambridge, Cambridge University Press.