Hindawi Publishing Corporation
Journal of Biomedicine and Biotechnology
Volume 2012, Article ID 450967, 8 pages
doi:10.1155/2012/450967

*Research Article*

# Bayesian Integration of Isotope Ratio for Geographic Sourcing of Castor Beans

**Bobbie-Jo Webb-Robertson,[1] Helen Kreuzer,[2] Garret Hart,[3] James Ehleringer,[4] Jason West,[5] Gary Gill,[6] and Douglas Duckworth[3]**

[1] *Computational Biology and Bioinformatics, Pacific Northwest National Laboratory, Richland, WA 99352, USA*
[2] *Biodefense, Pacific Northwest National Laboratory, Richland, WA 99352, USA*
[3] *Nuclear Material Analysis, Pacific Northwest National Laboratory, Richland, WA 99352, USA*
[4] *Department of Geology and Geophysics, The University of Utah, Salt Lake City, UT 84112, USA*
[5] *Department of Ecosystem Science and Management, Texas A&M University, College Station, TX 77843, USA*
[6] *Marine Sciences Laboratory, Pacific Northwest National Laboratory, Sequim, WA 98382, USA*

Correspondence should be addressed to Bobbie-Jo Webb-Robertson, bobbie-jo.webb-robertson@pnl.gov

Received 29 February 2012; Accepted 13 May 2012

Academic Editor: Carlos Ramos



Recent years have seen an increase in the forensic interest associated with the poison ricin, which is extracted from the seeds of the *Ricinus communis* plant. Both light element (C, N, O, and H) and strontium (Sr) isotope ratios have previously been used to associate organic material with geographic regions of origin. We present a Bayesian integration methodology that can more accurately predict the region of origin for a castor bean than individual models developed independently for light element stable isotopes or Sr isotope ratios. Our results demonstrate a clear improvement in the ability to correctly classify regions based on the integrated model with a class accuracy of 60.9 ± 2.1% versus 55.9 ± 2.1% and 40.2 ± 1.8% for the light element and strontium (Sr) isotope ratios, respectively. In addition, we show graphically the strengths and weaknesses of each dataset in respect to class prediction and how the integration of these datasets strengthens the overall model.

## 1. Introduction

Castor bean is the common term used for the seed of the plant *Ricinus communis*. Castor beans have a long history as a commercial crop through the world and thus are a valuable commercial commodity [1]. However, castor beans also contain the toxic protein ricin, which is classified as a Schedule 1 controlled substance under the Chemical Weapons Convention and as a select agent by several other agencies. In particular, very little ricin is necessary for a lethal dose, estimated to be only 5–10 micrograms per kilogram of body weight if injected or inhaled and 1–20 milligrams if ingested [2]. Although few deaths to date have been attributed to ricin poisoning [3], recent years have seen an increase in the seizure of ricin-containing samples related to biocriminal activity [4]. Thus, methods to track the source of castor beans have the potential to be of value to investigators. Recent methods associated with the attribution of ricin-containing samples have focused on characterization of the procedure that could have been used to extract ricin from the seeds [5, 6]. These methods give valuable information to the investigator. However, the information is associated with the extracted ricin and not the castor bean. Additionally, under the circumstance that the investigator collects castor beans in lieu of the processed product, it may be of interest to discover the region from which those castor beans originated.

By integrating data from a variety of analytical instruments, it may be possible to help identify the geographic source of castor seeds. These integration methods promise to yield a more complete and accurate view of a sample compared to any individual data source. However, heterogeneous data collected from different analytical methods cannot be simply concatenated together to build a statistical model. In addition, adequately orthogonal datasets are required to

construct an integration-based classifier that is more accurate than the individual models.

Here, we investigate two sources of isotope ratio (IR) data that may provide insight into the region of origin for castor beans: (1) light element (C, N, O, and H) stable isotope ratios (LeIRs) and (2) Sr isotope ratios (SrIRs). Both data types have been used to associate plant and animal material with regions of origin [7–11]. Stable isotope ratios of C, O, and H in plants are influenced by climate [12], while $^{87/86}$Sr isotope ratios are influenced by bedrock and soil [13], suggesting that they should be treated as independent datasets. We find that each dataset can predict region of origin more accurately than expected by chance, but with moderate success. Further, we present a statistical integration approach based on a simple Bayesian network to utilize the two datasets in combination to predict region of origin (Figure 1). The integrated classification model significantly improves our capability to predict region of origin. In addition, this statistical integration approach simultaneously yields a classification prediction as well as a probabilistic confidence in the identification.

## 2. Experimental Material and Methods

The castor seeds used for this study were a subset of a larger collection assembled by various means including purchasing seeds in various locales, accepting donations of seeds from collaborators, and sending seeds to volunteer growers throughout the United States [14]. The collection includes ornamental and agricultural varieties of castor seeds; our goal was to gather seeds from as geographically diverse regions as possible, and we accepted all acquisitions regardless of strain. Consequently, the collection is genetically variable. The actual growth conditions of the seeds were also not controlled; seeds could have been from irrigated plots, watered only by precipitation or any other method. From this broader collection, we subsampled seeds from 8 geographic regions from which we had multiple acquisitions to determine whether, despite the variation in genetics and potential cultivation method, there were characteristic isotope ratio values that linked the seeds to their regions of origin.

The nature of the seed collection imposed significant limitations in our definition of growth region. Ideally, a growth region would consist of an area homogeneous in geology, climate, and isotope ratios of precipitation and surface water from which we would have multiple seed acquisitions. Given the opportunistic nature of the castor seed collection, however, it lacked the sampling density in tightly constrained geographic regions required for the ideal experimental design. We therefore defined regions from diverse global locations in which we had multiple acquisitions in relatively limited areas. The sample set used in our analyses consisted of 68 castor seed acquisitions from 8 such geographic regions (Table 1).

We defined two categories of data: (1) $\delta^{13}$C, $\delta^{15}$N, $\delta^{18}$O, and $\delta^{2}$H isotope ratios (LeIRs) of the seeds and (2) $^{87}$Sr/$^{86}$Sr isotope ratios (SrIRs) of the seeds. The average value of the observations within each of these data categories is summarized in Table 1. Boxplots of the data are given in Figures 2 and 3 for the observed LeIRs and SrIR data, respectively.

Table 1: Sample sizes and average values for each of the 8 regions for the LeIR and SrIR data.

| Region | No. Obs | Avg. *IR* of seed | | | | Avg. |
|---|---|---|---|---|---|---|
| | | $\delta^{13}$C | $\delta^{15}$N | $\delta^{18}$O | $\delta^{2}$H | $^{87/86}$Sr |
| US01 (CA) | 7 | −26.26 | 5.97 | 23.27 | −121.89 | 0.710 |
| US02 (AZ) | 4 | −27.15 | 8.04 | 24.20 | −157.90 | 0.711 |
| US03 (UT) | 7 | −28.57 | 7.50 | 19.11 | −183.42 | 0.709 |
| US04 (TX) | 8 | −26.63 | 4.34 | 22.52 | −143.75 | 0.711 |
| BRAZ01 | 15 | −27.07 | 8.47 | 21.88 | −136.45 | 0.715 |
| BRAZ02 | 7 | −27.17 | 7.74 | 22.55 | −136.17 | 0.726 |
| CHIN | 9 | −27.62 | 5.98 | 17.72 | −171.66 | 0.711 |
| INDI | 11 | −27.85 | 8.22 | 23.72 | −138.34 | 0.715 |

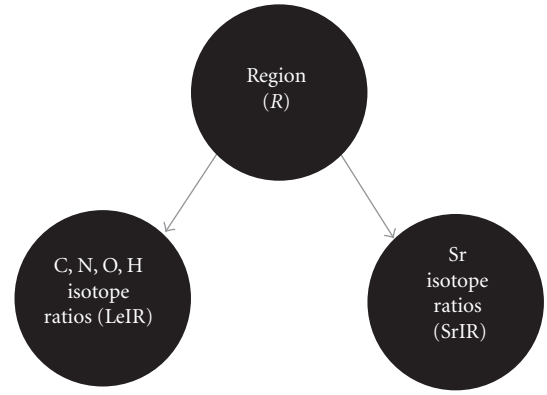


Figure 1: Basic Bayesian network formulation used for integration of the light element and Sr isotope ratios, LeIR and SrIR, respectively.

*2.1. Light Element (C, N, O, and H) Isotope Ratios.* Light element stable isotope ratios of castor beans were measured by isotope ratio mass spectrometry as described in [14, 15]. In brief, five castor beans from a single geographic source were homogenized using a Retsch MM200 machine (Retsch GmbH & Co., Germany). The C and N stable isotope ratios of the paste were determined on a Finnegan MAT Delta X isotope ratio mass spectrometer (Bremen, Germany) coupled to a Carlo Erba Elemental Analyzer 1108. The O and H stable isotope ratios of the paste were determined using a Thermo-Finnegan Delta Plus XL isotope ratio mass spectrometer (Bremen, Germany) equipped with a thermal conversion elemental analyzer (TC/EA). Stable isotope content is measured as a ratio, R (e.g., $^{13}$C/$^{12}$C), and reported as a delta ($\delta$) value where $\delta = [(R_{\text{sample}}/R_{\text{standard}}) - 1]^{*}1,000\%$. In this equation, $R_{\text{sample}}$ is the measured isotope ratio of the sample, and $R_{\text{standard}}$ is the isotope ratio of an internationally recognized standard. The standard for C isotope ratio measurement is Vienna PeeDee Belemnite (VPDB), for N is air (AIR), and for O and H is Vienna Standard Mean Ocean Water (VSMOW) [16].

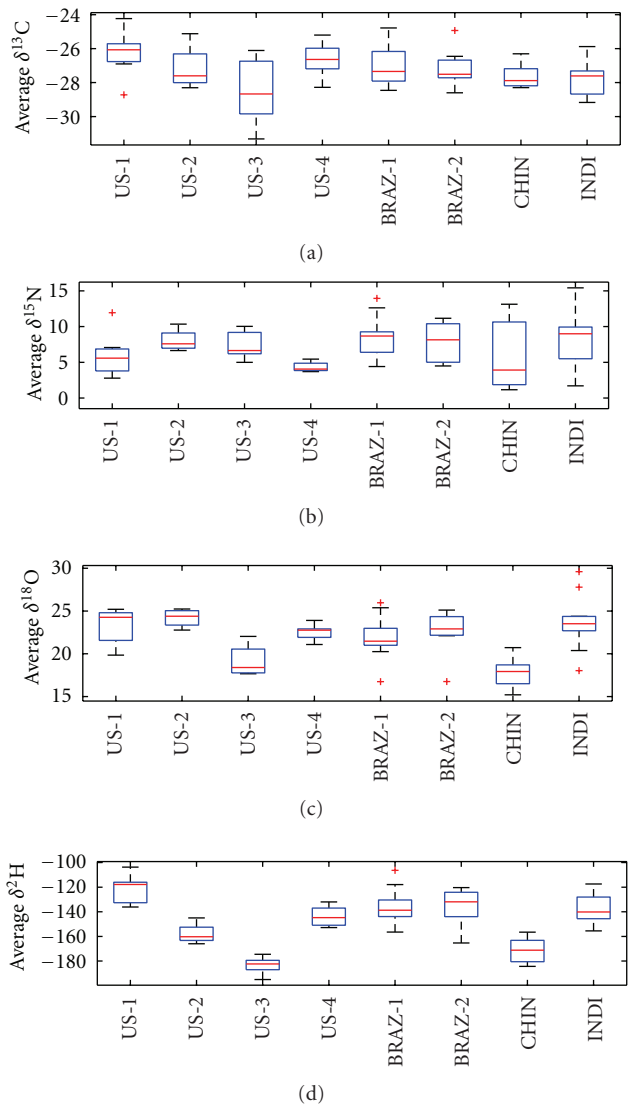


Figure 2: Boxplots showing the spread and deviation within each region for each light element IR.

*2.2. Sr Isotope Ratios.* Sr isotopes were measured by digesting 1-2 castor bean with 2-3 treatments of concentrated nitric acid and hydrogen peroxide coupled with heating and drying in order to break down all the organics. Strontium was separated from the digested sample using Sr-Spec (Eichrom) resin and nitric acid. The eluted Sr sample was dried and treated with 1-2 drops concentrated acid to further drive off organics and then reconstituted to a final volume of 4 mL of 2% nitric acid. The Sr isotope analyses were performed on a MC-ICPMS (Neptunen *Plus*) using a standard spray chamber and a self-aspirating nebulizer on 50 ppb solutions. For quality control, NBS-987 was run along with the unknowns ($^{87}$Sr/$^{86}$Sr = 0.71026 ± 1; $n$ = 5). The analyses were corrected for mass bias using $^{86}$Sr/$^{88}$Sr = 0.1194 and normalized to a NBS-987 standard value of 0.71024.

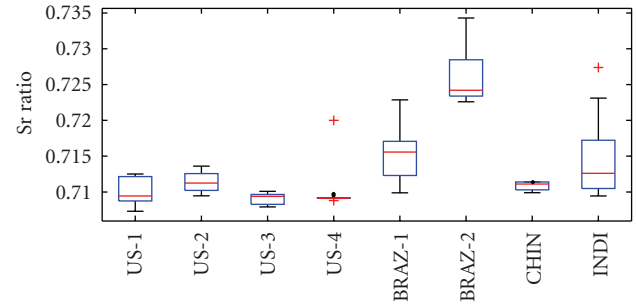


Figure 3: Boxplot showing the spread and deviation within each region for the SrIR measurements.

## 3. Statistical Material and Methods

The statistical model is formulated as a simple Bayesian network (Figure 1). Bayesian statistics is a common approach to make inferences from biological data because all data are treated as random variables. Bayesian models provide a full joint distribution over both the observable and unobservable variables (1). Furthermore, the posterior probability of interest can be computed by integration or summation, such as viewed in (2) [17, 18]. In particular, for the Bayesian formulation in Figure 1, the random variable region ($R$) is conditionally dependent upon each data type; however, the sources of data are not conditionally related to each other. Thus, the joint probability can be described by these conditional relationships

$$\text{Joint} = P(\text{LeIR}, \text{SrIR}, R) = P(\text{LeIR} \mid R)\ P(\text{SrIR} \mid R)P(R). \tag{1}$$

The specific probability of interest is the probability of observing region $k$ given our two data types, which can be obtained directly by applying Bayes formula to (1),

$$P(R_k \mid \text{LeIR}, \text{SrIR}) = \frac{P(\text{LeIR}, \text{SrIR}, R_k)}{P(\text{LeIR}, \text{SrIR})} = \frac{P(\text{LeIR} \mid R_k)P(\text{SrIR} \mid R_k)P(R_k)}{\sum_k P(\text{LeIR} \mid R_k)P(\text{SrIR} \mid R_k)P(R_k)}. \tag{2}$$

Thus, the task of computing the probability of interest in (2) simplifies to computing the posterior probability models of $P(\text{LeIR}|R)$ and $P(\text{SrIR}|R)$.

*3.1. Individual Posterior Probability Models.* MATLAB 2011b with Statistics Toolbox V7.6 was used to perform all statistical analyses on the LeIR and SrIR datasets, as well as the integration and validation of the models.

*3.1.1. Light Element (C, N, O, and H) Isotope Ratios.* The light element stable isotope ratios (IRs) consisted of four variables with little colinearity. These variables were relatively normally distributed with $P$ values ranging from 0.24 to 0.5 based on a Jarque-Bera test of normality [19]. Boxplots of the distribution of each variable are given in Figure 2. Given the

normal structure of the data and a categorically distributed dependent variable (regions), linear discriminant analysis (LDA) was used to derive a statistical classification model. LDA is a multivariate discrimination method commonly used for classification in chemometrics [20]. LDA uses statistical learning to infer an optimal linear combination of the features to separate the regions. The "classify" function in MATLAB was used to obtain the probability of region $k(R_k)$ given a set of IR values. The statistical model based on training data can be described as

$$P(R_k \mid \delta^{13}\mathrm{C}, \delta^{15}\mathrm{N}, \delta^{18}\mathrm{O}, \delta^{2}\mathrm{H}) = f_k(\mathrm{LeIR}), \quad (3)$$

where $f_k(\mathrm{LeIR})$ is computed from a multivariate normal distribution. The posterior probability of interest (2) for a test sample $j(\mathrm{LeIR}_j)$ is computed directly from the "classify" function

$$P(\delta^{13}\mathrm{C}, \delta^{15}\mathrm{N}, \delta^{18}\mathrm{O}, \delta^{2}\mathrm{H} \mid R_k) = P\left(\mathrm{LeIR}_j \mid R_k\right). \quad (4)$$

*3.1.2. Sr Isotope Ratios.* The SrIR data is described by a single observed value, and unlike the LeIR, the observed data for Sr is not normally distributed ($P$ value of approximately 0.001). A boxplot of the distribution of Sr across regions is given in Figure 3. Given that the data is non-normal with a single independent variable and categorically distributed dependent variable, multinomial logistic regression (MLR) was used to derive a statistical classification model [21] using the "mrnfit" function in MatLab

$$P(R_k \mid \mathrm{Sr}) = \frac{e^{(\mathrm{Sr}\bullet\beta_k)}}{1 + \sum_k e^{(\mathrm{Sr}\bullet\beta_k)}}, \quad (5)$$

where $\beta_k$ is the vector of regression coefficients for region $k$. Again the posterior is computed in MATLAB using the "mrnval" function to obtain $P(\mathrm{SrIR} \mid R_k)$.

*3.2. Classification Model Evaluation Metrics.* Each model is evaluated independently using a leave-one-out bootstrapping cross-validation approach (LOOB-CV) with resampling [22] to obtain the full set of posterior probabilities for each sample in our datasets. The LOOB-CV method was selected to reduce the likelihood of overtraining the model, and resampling is performed to acquire uncertainty estimates on the metrics of model accuracy. In particular, for the LOOB-CV method, each sample is left out to create $N$ datasets $[X_{-1}, X_{-2}, \ldots, X_{-N}]$, where $N = 68$ for the data described in Section 2. A set of 100 bootstrap samples, each containing 50 samples, are randomly selected for each $X_{-i}$, $[B_{-i}^{(1)}, B_{-i}^{(2)}, \ldots, B_{-i}^{(100)}]$. The model is trained on each $B_{-i}^{(k)}$, and the posterior of sample $i$ is obtained. The posteriors across the 100 bootstrap samples are averaged to obtain a more accurate estimate of the posterior probability.

The results are evaluated using two approaches: (1) average classification accuracy (CA) and (2) average area under a receiver operating characteristic curve (AUC). To compute these, each sample was defined by a binary vector where all values are initialized to zero. The probabilities for the sample were sorted, and all locations equal to or greater than the correct answer were set to 1. For example, suppose the correct region has the third largest probability of the 8, then it is set to $S_i = [0, 0, 1, 1, 1, 1, 1, 1]$. If the correct region is identified as the most probable then this becomes $S_i = [1, 1, 1, 1, 1, 1, 1, 1]$. The CA is defined as the fraction of samples that are correctly classified into the appropriate region

$$\mathrm{CA} = \frac{\sum_i S_{i1}}{N}. \quad (6)$$

The AUC is computed based on a modified receiver operating characteristic (ROC) curve. ROC curves traditionally plot the false positive rate (FPR) versus true positive rate (TPR) of a binary classifier. For this data, there are 8 possible classes with one correct answer and seven potential false identifications. Thus, there may be 0, 1, 2, 3, 4, 5, 6 or 7 regions incorrectly identified prior to correct answer, which means that each sample may have an FPR defined as one of these eight values [0, 1/7, 2/7, 3/7, 4/7, 5/7, 6/7, 1]. Based on these FPR states, the associated TPR for the full dataset is computed as the total number of samples that have an FPR less than or equal to the defined value,

$$tpr_m = \frac{\sum_i S_{im}}{N}. \quad (7)$$

Note that $\mathrm{CA} = tpr_1$. Similar to a binary ROC curve, a perfect classifier would identify all samples correctly with no regions correctly identified ahead of the true classification, which would result in an AUC of 1.0. Additionally, a random permutation would yield a linear relationship between the FPR and the TPR and an AUC of 0.5.

## 4. Results and Discussion

Prior to evaluation of the improvement in classification accuracy based on the integrated model, an exhaustive evaluation of the variables that could be used for the LeIR data model is performed. There are 15 possible combinations of the four light element IRs (Table 2). For each combination, the average CA and the average AUC are computed (6) and (7) and are ordered in Table 2 by average CA. The most accurate model is the one that includes all variables, although the models excluding either O, H, or both are very similar. The average ROC, however, is nearly identical for these top four models. This result is consistent with past observations that climatic factors and thus geography influence the C, O, and H isotope ratios of plants [23–28]. These effects are likely to be similar for plants growing in similar geographic locales. N isotope ratios of a single species, in contrast, are a function of N sources such as fertilizer (if any) and are likely to be independent of geography, at least in most cases [29]. However, since the inclusion of N improves our ability to classify these regions and the IR of O may be highly valuable to regions not included in our sample dataset and it does not decrease accuracy, we utilize the LDA model that included all four variables for the development of statistical models for the purposes of integration.

The LOOB-CV analysis was performed 100 times, which yielded an average CA of 55.9% ± 2.1% and 40.2% ± 1.8%

Table 2: The average CA and AUC for all possible combination of the variables in the LeIR data.

| Variables | | | | Average CA | Average AUC |
|---|---|---|---|---|---|
| $\delta^2$H | $\delta^{18}$O | $\delta^{13}$C | $\delta^{15}$N | 55.9% | 0.88 |
| $\delta^2$H | | $\delta^{13}$C | $\delta^{15}$N | 55.2% | 0.87 |
| | $\delta^{18}$O | $\delta^{13}$C | $\delta^{15}$N | 53.3% | 0.88 |
| | | $\delta^{13}$C | $\delta^{15}$N | 52.2% | 0.87 |
| $\delta^2$H | $\delta^{18}$O | | $\delta^{15}$N | 45.1% | 0.84 |
| | $\delta^{18}$O | | $\delta^{15}$N | 42.8% | 0.83 |
| $\delta^2$H | | | $\delta^{15}$N | 36.8% | 0.83 |
| $\delta^2$H | | $\delta^{13}$C | | 35.8% | 0.78 |
| $\delta^2$H | $\delta^{18}$O | $\delta^{13}$C | | 35.0% | 0.78 |
| | $\delta^{18}$O | $\delta^{13}$C | | 33.9% | 0.78 |
| | | $\delta^{13}$C | | 31.4% | 0.75 |
| | | | $\delta^{15}$N | 30.7% | 0.82 |
| $\delta^2$H | $\delta^{18}$O | | | 23.5% | 0.67 |
| | $\delta^{18}$O | | | 23.0% | 0.63 |
| $\delta^2$H | | | | 19.1% | 0.69 |

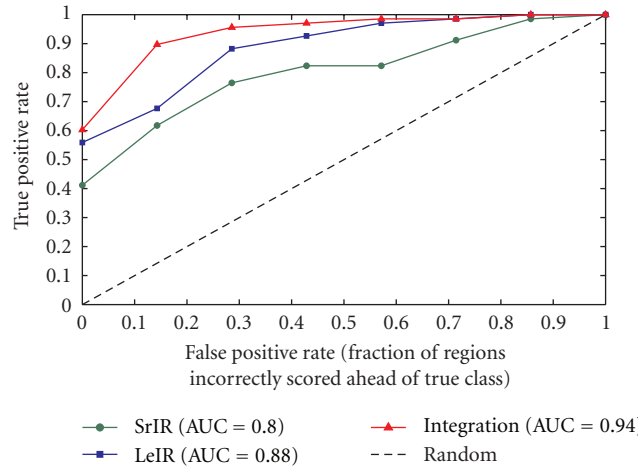


Figure 4: Modified ROC curves to evaluate the overall capability of each data type to predict region versus the integrated model.

for the LeIR and SrIR datasets, respectively. The integrated posterior probability was computed as described in (2), and the overall classification accuracy improved to 60.9% ± 2.1%. This is a significant improvement over the individual IR model; the null hypothesis that the average difference in the means between the two models is equal to zero is rejected with a $P$ value less than $1e-40$ (based on a two-sample $t$-test). The ROC curve (Figure 4), shows this is a clear trend with an improvement of the AUC to 0.94 versus 0.88 and 0.80 for LeIR and SrIR, respectively. The 100 sampled observations of the AUC of the integrated model are significant larger than the LeIR model at a $P$ value $< 1e-10$ (based on a two-sample $t$-test), and the curves depicted in Figure 4 are significantly different at a $P$ value of 0.06 (based on a sign rank nonparametric paired test) [30]. Thus, although SrIR does not perform well alone, it does offer a significant contribution if integrated with the LeIR data.

The evaluation of the datasets via the CA and AUC gives an overall view of the predictability of the datasets with respect to region but does not give insight across the regions. The improved accuracy of the integrated model points to a capability of one dataset to correctly predict the appropriate region with higher probability than the misclassified probability of the other dataset. To evaluate the classification accuracy in respect to region, a visualization akin to that used in Visual Integration for Bayesian Evaluation (VIBE) was employed [31]. Figure 5 gives classification accuracy plot across the 8 regions as the true class on the $y$-axis and predicted class on the $x$-axis. A perfect classifier would have a diagonal of solid black since all of the predicted classes would be equal to the true classes and have a value of 1. The classification accuracies observed in Figure 5 are slightly different than above since this is a single sample from the 100 iterations performed. However, a similar trend is observed in which the IR data has a larger CA than the Sr data and the integrated model outperforms either alone. The class accuracy plot quickly demonstrates that the IR data has the most challenge distinguishing regions 5 through 8 (outside of the united states), and most often regions are misclassified into region 4 (US04 = TX)—see Table 1. The Sr data has a completely different profile in terms of how it correctly and incorrectly classifies samples. It most often misclassifies samples into either the third or fifth regions, US03 (UT) and BRAZ01. In addition, it never correctly classifies any samples from regions 1, 2, 4, and 8. The integrated model on the right corrects many of the imbalances observed from the individual datasets. The specific geographic regions that are being correctly classified are easily distinguished from the class accuracy plot, as well as those regions that cannot be easily distinguished by isotope ratios.

The relative lack of power of the O and H isotope ratio data to link seeds to their regions of origin initially appeared surprising, as O and H isotope ratios of plant material have been shown to be linked to geographic region of origin [12]. Plants derive O and H atoms from their water sources, while there is a strong and well-recognized link between the isotope ratios of precipitation and geography [32]. However, the limitations on experimental design imposed by nature of the seed collection might lead one to expect this effect. Some of the defined growth regions spanned a gradient of climate. For example, the "Texas" region included samples from the vicinity of Lubbock, which has an arid climate, and Houston, which is quite humid. The surface water isotope ratios of these two parts of the state are predicted to differ somewhat, based on US Geological Survey data [33]. Thus, the source water accessed the plants as well as the extent of evaporative enrichment of plant leaf water, a source used for biosynthesis of many plant organic components [12], would likely be expected to differ and result in differing O and H isotope ratios in the seeds. A higher sampling density in strictly limited geographic regions would permit a better analysis of the effect of O and H isotope ratios on region of origin association. Another possible reason for the lack of power of O and H isotope ratios to associate seeds with

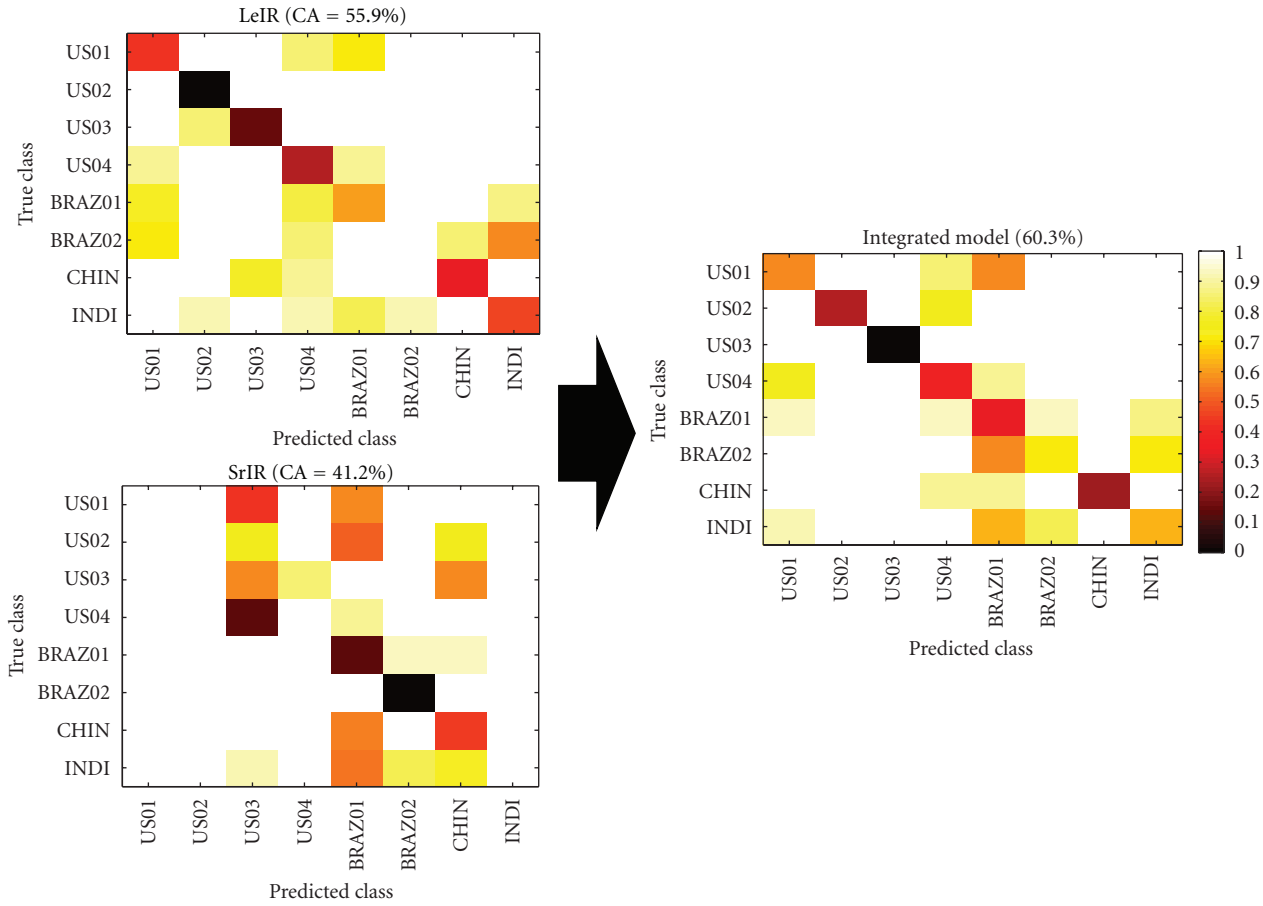


Figure 5: Class accuracy plots show what fractions of the samples are being classified into specific groups, allowing a direct comparison of the true versus predicted classes: the integrated model on the right shows a clear improvement in overall matches between the true and predicted classes.

regions of origin is the potential variation in cultivation conditions of the seeds. Even if a region could be more tightly defined, cultivation practices could influence the actual isotope ratios of the water used by the castor plants. For example, irrigation via open ditches could result in significant evaporative enrichment in isotopic content of the water taken up by the plants, while drawing water from a deep well could provide them with water isotopically different from surface water. Controlling the environmental variation imposed by cultivation conditions might also improve the discrimination power of the O and H isotope ratios. Finally, the genetic variability of the castor seeds may affect water dynamics within the individual plants, which could impart some variation to the O and H isotope ratios of plants growing in identical environments.

The ability of isotope ratios to associate castor seeds with region of origin despite the limitations imposed by sampling density, the genetic variability of the seeds, and the probable variation in cultivation methods is noteworthy. Presumably, if we had been able to more strictly define geographic regions to those with homogenous climate and geology and control the variables of genetics and growth conditions, the accuracy of the association would be improved, perhaps significantly. Further experiments with multiple seed acquisitions from tightly defined source regions could address this question. For real-world application of isotope ratios for assigning region of origin, however, it is unlikely that either the genetic strain of the sample (whether of castor seeds or some other plant product such as a food) or its cultivation method would be controllable. Demonstrating that integrated isotopic data can associate a plant product with its region of origin in the absence of such control suggests that this approach could be broadly useful for geographic sourcing.

## 5. Conclusion

Both light element (C, N, O, and H) stable isotope ratios and $^{87/86}$Sr isotope ratios have been used to associate plant and animal materials with its geographic region of origin. Here, we show that each of these datasets independently can associate castor seeds with region of origin more accurately than would be expected by chance, as shown by both classification accuracy and a modified ROC curve model. Bayesian integration of these two data streams yielded results that were significantly better than those from either individual dataset.

This approach illustrates the benefits afforded by a rigorous approach to data integration and its application to forensics community.

## Acknowledgments

This work was supported in part by Laboratory Directed Research and Development at Pacific Northwest National Laboratory (PNNL) and the National Science Foundation (Grant 0743543). PNNL is a multiprogram national laboratory operated by Battelle for the U.S. Department of Energy under Contract DE-AC06-76RL01830.

## References

[1] E. A. Weiss, *Castor. Oilseed Crops*, Longman, London, UK, 1983.

[2] S. M. Bradberry, K. J. Dickers, P. Rice, G. D. Griffiths, and J. A. Vale, "Ricin poisoning," *Toxicological Reviews*, vol. 22, no. 1, pp. 65–70, 2003.

[3] K. R. Challoner and M. M. McCarron, "Castor bean intoxication," *Annals of Emergency Medicine*, vol. 19, no. 10, pp. 1177–1183, 1990.

[4] J. M. Bale et al., *Ricin Found in London: An al-Qa'ida Connection?* James Martin Center for Nonproliferation Studies, Monterey, Calif, USA, 2002.

[5] H. A. Colburn, D. S. Wunschel, H. W. Kreuzer, J. J. Moran, K. C. Antolick, and A. M. Melville, "Analysis of carbohydrate and fatty acid marker abundance in ricin toxin preparations for forensic information," *Analytical Chemistry*, vol. 82, no. 14, pp. 6040–6047, 2010.

[6] H. W. Kreuzer, J. H. Wahl, C. N. Metoyer, H. A. Colburn, and K. L. Wahl, "Detection of acetone processing of castor bean mash for forensic investigation of ricin preparation methods," *Journal of Forensic Sciences*, vol. 55, no. 4, pp. 908–914, 2010.

[7] T. E. Cerling, G. Wittemyer, H. B. Rasmussen et al., "Stable isotopes in elephant hair document migration patterns and diet changes," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 2, pp. 371–373, 2006.

[8] J. R. Ehleringer, J. F. Casale, M. J. Lott, and V. L. Ford, "Tracing the geographical origin of cocaine: Cocaine carries a chemical fingerprint from the region where the coca was grown," *Nature*, vol. 408, no. 6810, pp. 311–312, 2000.

[9] J. R. Ehleringer, D. A. Cooper, M. J. Lott, and C. S. Cook, "Geo-location of heroin and cocaine by stable isotope ratios," *Forensic Science International*, vol. 106, no. 1, pp. 27–35, 1999.

[10] S. Swoboda, M. Brunner, S. F. Boulyga, P. Galler, M. Horacek, and T. Prohaska, "Identification of Marchfeld asparagus using Sr isotope ratio measurements by MC-ICP-MS," *Analytical and Bioanalytical Chemistry*, vol. 390, no. 2, pp. 487–494, 2008.

[11] J. B. West, J. R. Ehleringer, and T. E. Cerling, "Geography and vintage predicted by a novel GIS model of wine $\delta^{18}O$," *Journal of Agricultural and Food Chemistry*, vol. 55, no. 17, pp. 7075–7083, 2007.

[12] J. B. West, H. W. Kreuzer, J. R. Ehleringer et al., "Approaches to plant hydrogen and oxygen isoscapes generation," in *Isoscapes: Understanding Movement, Pattern, and Process on Earth through Isotope Mapping*, J. B. West, Ed., pp. 161–178, Springer, Monterey, Calif, USA, 2010.

[13] B. L. Beard and C. M. Johnson, "Strontium isotope composition of skeletal material can determine the birth place and geographic mobility of humans and animals," *Journal of Forensic Sciences*, vol. 45, no. 5, pp. 1049–1061, 2000.

[14] H. W. Kreuzer, J. B. West, and J. R. Ehleringer, "Forensic applications of light-element stable isotope ratios of Ricinus communis seeds and ricin preparations ," *Journal of Forensic Sciences*. In press.

[15] J. R. Ehleringer et al., "Stable isotope ratio analyses of castor bean: a ricin signature program," in *Federal Bureau of Investigation*, 2006.

[16] T. B. Coplen, "New guidelines for reporting stable hydrogen, carbon and oxygen isotope-ratio data," *Geochimica et Cosmochimica Acta*, vol. 60, no. 17, pp. 3359–3360, 1996.

[17] J. M. Bernardo and A. F. M. Smith, *Bayesian Theory*, John Wiley & Sons, New York, 2000.

[18] B. J. Webb-Robertson, L. A. McCue, N. Beagley et al., "A Bayesian integration model of high-throughput proteomics and metabolomics data for improved early detection of microbial infections," *Pacific Symposium on Biocomputing*, pp. 451–463, 2009.

[19] C. M. Jarque and A. K. Bera, "A test for normality of observations and regression residuals," *International Statistical Review*, vol. 55, no. 2, pp. 163–172, 1987.

[20] K. R. Beebe, R. J. Pell, and M. B. Seasholtz, *Chemometrics: A Practical Guide*, John Wiley & Sons, Hoboken, NJ, USA, 1998.

[21] P. McCullagh and J. A. Nelder, *Generalized Linear Models*, Chapman & Hall, New York, NY, USA, 1990.

[22] W. Jiang and R. Simon, "A comparison of bootstrap methods and an adjusted bootstrap approach for estimating the prediction error in microarray classification," *Statistics in Medicine*, vol. 26, no. 29, pp. 5320–5334, 2007.

[23] G. D. Farquhar, J. R. Ehleringer, and K. T. Hubick, "Carbon isotope discrimination and photosynthesis," *Annual Review of Plant Physiology and Molecular Biology*, vol. 40, pp. 503–537, 1989.

[24] G. D. Farquhar and R. A. Richards, "Isotopic composition of plant carbon correlates with water-use efficiency of wheat genotypes," *Australian Journal of Plant Physiology*, vol. 11, no. 6, pp. 539–552, 1984.

[25] G. D. Farquhar, M. H. O'Leary, and J. A. Berry, "On the relationship between carbon isotope discrimination and the intercellular carbon dioxide concentration in leaves," *Australian Journal of Plant Physiology*, vol. 9, no. 2, pp. 121–137, 1982.

[26] L. Sternberg, M. DeNiro, and R. Savidge, "Oxygen isotope exchange between metabolites and water during biochemical reactions leading to cellulose synthesis," *Plant Physiology*, vol. 82, pp. 423–427, 1986.

[27] J. Gray and P. Thompson, "Climatic information from $^{18}O/^{16}O$ ratios of cellulose in tree rings," *Nature*, vol. 262, no. 5568, pp. 481–482, 1976.

[28] S. Epstein, P. Thompson, and C. J. Yapp, "Oxygen and hydrogen isotopic ratios in plant cellulose," *Science*, vol. 198, no. 4323, pp. 1209–1215, 1977.

[29] L. H. Pardo and K. J. Nadelhoffer et al., "Using nitrogen isotope ratios to assess terrestrial ecosystems at regional and global scales," in *Isoscapes: Understanding Movement, Pattern, and Process on Earth through Isotope Mapping*, J. B. West et al., Ed., pp. 221–250, Springer, Dordrecht, The Netherlands, 2010.

[30] R. L. Ott and M. Longnecker, *An Introduction to Statistical Methods and Data Analysis*, Brooks/Cole, Belmont, 6th edition, 2010.
[31] N. Beagley, K. G. Stratton, and B. J. Webb-Robertson, "VIBE 2.0: visual integration for bayesian evaluation," *Bioinformatics (Oxford, England)*, vol. 26, no. 2, pp. 280–282, 2010.
[32] H. Craig, "Isotopic variations in meteoric waters," *Science*, vol. 133, no. 3465, pp. 1702–1703, 1961.
[33] C. Kendall and T. B. Coplen, "Distribution of oxygen-18 and deuteriun in river waters across the United States," *Hydrological Processes*, vol. 15, no. 7, pp. 1363–1393, 2001.