DNA Pool Analysis-based Forgery-Detection of Dairy Products

International Journal of Electrical and Computer Engineering (IJECE)
Vol. 8, No. 5, October 2018, pp. 3913 – 3922
ISSN: 2088-8708, DOI: 10.11591/ijece.v8i5.pp3913-3922

DNA Pool Analysis-based Forgery-Detection of Dairy Products

Francesco Rossi[1], Paola Modesto[2], Cristina Biolatti[3], Alfredo Benso[4], Stefano Di Carlo[5], Gianfranco Politano[6], and Pierluigi Acutis[7]
[1,4,5,6]Department of Control and Computer Engineering, Politecnico di Torino, Italy
[2,3,7]Istituto Zooprofilattico Sperimentale del Piemonte Liguria e Valle dAosta, Italy

Article Info

Article history:
Received December 21, 2017
Revised July 29, 2018
Accepted August 18, 2018

Keyword:
Genetic Programming
CMA-ES
DNA barcoding
STR
Food Safety

ABSTRACT

Food integrity and food safety have received much attention in recent years due to the dramatic increasing number of food frauds. In this article we focus on the problem of dairy products traceability. In particular, we propose an automatic forgery detection system able to detect frauds in milk and cheese. We investigate the use of Short Tandem Repeats analysis data, processed by a Covariance Matrix Adaptation Evolution Strategy algorithm in order to evaluate a traceability score between the products and their producer, and to highlight possible adulterations and inconsistencies. To demonstrate the usability of the proposed heuristic algorithm in a real setup, we also present the results collected from two real Italian farms.

Corresponding Author:
Francesco Rossi
Politecnico di Torino
Department of Control and Computer Engineering
Corso Duca delgli Abruzzi 24, 10129 Torino, Italy
Email: francesco.rossi@polito.it

1. INTRODUCTION

Food integrity and food safety have received much attention in recent years due to the dramatic increasing number of food frauds. Traceability is a useful method to guarantee foodstuff quality and safety, to guarantee hygiene standards, and to protect consumers choices and health. Over the past years DNA analysis has been widely recognized as an effective tool to deal with genetic traceability issues, gaining a key role in tracing and testing food origin and safety.

In this article we analyze dairy products for which one of the crucial issues is traditional cheese traceability. In the case of frauds, it may occur that a selected dairy product that should be produced by milk coming from a certified farm, is instead produced using a variable amount of milk coming from unauthorized farms. Traceability of dairy products through DNA analysis involves some technical challenges. The cheese (CH) is produced from bulk milk (BM), which contains DNA from different cows of the farm and undergoes several biochemical changes during the ripening process.

In this paper, we propose a computer-assisted molecular traceability system able to analyze the origin of a traditional dairy product. We investigate the use of Short Tandem Repeats (STRs) analysis to create a DNA fingerprint of small dairy farms and to link dairy products (milk and cheese) to the corresponding producer. So far, STR analysis has been applied to blood samples for genetics population analysis [1, 2, 3, 4, 5], or to milk samples in order to identify quantitative trait locus (QTL) associated with traits in animal science [6, 7]. However, the application of STR analysis to trace the origin of dairy products is a different and more complex issue. Dairy products contain the DNA belonging to several different individuals, preventing the possibility to perform single-animal traceability. In literature, dairy products traceability has been mainly addressed by studying Fatty Acids and Triacylglycerols Content using Gas Chromatography [8]. So far the STR marker analysis proved to be valid only in mono-breed setup to detect adulteration in dairy product [9].

To the best of our knowledge, this work is the first attempt to explore the use of pooled STR analysis for traceability of food products.

Two farms owning different cow breeds were included in this study. First, the DNA of each animal was analyzed to compute a DNA signature based on the analysis of known STRs loci. The same STR analysis was then performed on the final dairy products. The obtained STR genetic datasets were analyzed through a Covariance Matrix Adaptation Evolution Strategy (CMA-ES) algorithm in order to evaluate the correlation (and therefore traceability) between the dairy products and the corresponding set of animals that contributed to their production. As an outcome, the proposed algorithm was able to highlight possible adulterations and/or inconsistencies.

Results showed that bulk milk and derived cheese present an STR profile composed of a subgroup of the STRs identified in the animals the dairy product originated from, and the profile could be efficiently used to trace the origin of the dairy product.

2. RESEARCH METHOD

In this section, we describe the procedure followed to generate the STR datasets, and we present the proposed Computer-Assisted Molecular Traceability system and its implementation based on the CMA-ES [10] algorithm available in R [11].

2.1. STR Dataset

Two farms with different geographic locations and breed cows were considered for the tuning of the method. At the beginning of the study, appointed veterinaries collected blood and milk samples from each cow. Afterwards, they monthly sampled BM and CH for 12 months in the first farm and 11 months in the second one. All collected samples were cold-stored for the tuning of the analysis protocol and the choice of the best genotyping process. The main steps of the STRs selection and data generation can be summarized as follows:

- Sample Collection: DNA extraction from blood, milk somatic cells and cheese collected during the months;
- STRs selection: from a panel of 280 available STRs (from literature), 20 STRs were chosen taking into account some of their characteristics, as well as other technical parameters related to the tuning phase of the analysis protocol (the STR selection process is proprietary and, at the moment, it cannot be fully disclosed);
- Genotyping Process: capillary electrophoresis using a 3130 Genetic Analyzer (Applied Biosystems) and fragments sizing using the STRAnd software [12];
- Data extraction: the peak height of each allele in relative fluorescence unit (RFU) of the electropherogram track was considered as an indication of its quantity and used in the following analyses.

Once the genotyping process was completed, the obtained raw data were organized in a tabular format (Table 1) reporting the allele frequencies for each STR and for each cow. The notation in Table 1 must be read as follows:

- n is the number of processed STRs;
- m is the number of cows available within the examined farm;
- $a^{(i,j)}$ $(i \in [1,m], j \in [1,n])$ is the specific alleles dimension (bp) of the i^{th} cow for the j^{th} STR. This notation includes the indication of the polymorphism occurrence of being heterozygote ($a^{(i,j)x} \neq a^{(i,j)y}$) or homozygote ($a^{(i,j)x} = a^{(i,j)y}$).

Similarly, also the BM and the CH genotyping pool analysis data were organized in a tabular way (Table 2). However, differently from Table 1, the information associated to each cell aPj (PBM,CH, j[1,n]) of the table, is a vector including all the allele values obtained from the genotyping process of the pool P for the j^{th} STR.

Finally, the absolute RFU alleles peak (h) of each allele for each cow of the farm, for BM and for CH were organized according to Table 3. At the end all tabular data were stored in comma-separated values (CSV) format text files.

Table 1. Example of a data farm organization.Here the $a^{(i,j)x}$,$a^{(i,j)y}$ notation represents the two alleles for each cow in each STR.

Cows	STR1	STR2	STR3	...	STR*n*
COW1	$a^{(1,1)x}$,$a^{(1,1)y}$	$a^{(1,2)x}$,$a^{(1,2)y}$	$a^{(1,3)x}$,$a^{(1,3)y}$	...	$a^{(1,n)x}$,$a^{(1,n)y}$
COW2	$a^{(2,1)x}$,$a^{(2,1)y}$	$a^{(2,2)x}$,$a^{(2,2)y}$	$a^{(2,3)x}$,$a^{(2,3)y}$	...	$a^{(2,n)x}$,$a^{(2,n)y}$
COW3	$a^{(3,1)x}$,$a^{(3,1)y}$	$a^{(3,2)x}$,$a^{(3,2)y}$	$a^{(3,3)x}$,$a^{(3,3)y}$	...	$a^{(3,n)x}$,$a^{(3,n)y}$
...	...	...	...	...	...
COW*m*	$a^{(m,1)x}$,$a^{(m,1)y}$	$a^{(m,2)x}$,$a^{(m,2)y}$	$a^{(m,3)x}$,$a^{(m,3)y}$	...	$a^{(m,n)x}$,$a^{(m,n)y}$

Table 2. BM and CH data organization. Here $\underline{a_P^j}$ represents the pool P allele vector for each STR.

Pool	STR1	STR2	STR3	...	STR*n*
BM	$\underline{a_{BM}^1}$	$\underline{a_{BM}^2}$	$\underline{a_{BM}^3}$	...	$\underline{a_{BM}^n}$
CH	$\underline{a_{CH}^1}$	$\underline{a_{CH}^2}$	$\underline{a_{CH}^3}$	...	$\underline{a_{CH}^n}$

2.2. Computer-Assisted Molecular Traceability

The first experiments we performed attempted to evaluate the ability to trace dairy products using well known software algorithms commonly used in genetic distance analysis like FSTAT [13], PHYLIP [14] and SMOGD [15] and then resorting to STRUCTURE [16]. However, results showed that these algorithms were not well suited to accomplish the intended purpose. They usually apply a Bayesian algorithm approach to assign a sample genotype to a specific dataset representing the candidate group of origin. While they work well in diploid data (i.e. only two alleles), they did not perform properly in the experimental setup considered in this paper due to the presence of variable numbers of alleles for each STR in every sample (e.g. milk and cheese pooled DNA samples).

Therefore, we decided to implement a new approach able to detect if the BM or CH fingerprint could be traced and compared with the genetic pool characteristics of the producing farm. Our innovative method is at first glance an automatic heuristic procedure based on the Covariance Matrix Adaptation Evolution Strategy (CMA-ES) algorithm. The heuristic is employed to estimate the likelihood of an STRs profile of BM or CH to be originated by a combination of the STR profiles of the cows from which the dairy product was originated from.

The next subsection provides the reader with the general principles about the CMA-ES, which is necessary to better understand the proposed computer-assisted molecular traceability method described next.

2.2.1. CMA-ES algorithm

The covariance matrix adaptation evolution strategy (CMA-ES) is an optimization method first proposed by Hansen, Oster Meier, and Gawelczyk [17] and further developed in subsequent years [18, 19]. The CMA-ES performs an exploration in a solution space exploiting a covariance matrix, closely related to the inverse Hessian on convex-quadratic functions. The approach is particularly suited to solve difficult non-linear, non-convex, and non-separable problems, of at least moderate dimensionality (i.e. $n \in [10, 100]$).

In CMA-ES, iteration steps are called generations due to its biological foundations. The value of a generic algorithm parameter y during generation g is denoted with $y^{(g)}$. The mean vector $m^{(g)} \in R^n$ represents the favorite, most promising solution so far. The step size $\sigma^{(g)} \in R^+$ controls the step length, and the covariance matrix $C^{(g)} \in R^{n \times n}$ determines the shape of the distribution ellipsoid in the search space. Conversely, its goal is to fit the search distribution to the contour lines of the objective function f to be minimized: $C^{(0)} = I$.

One of the main characteristics of the CMA-ES is that it requires almost no parameter tuning for its application unlike most common heuristic optimization methods [20]. The choice of its internal parameters is not left to the user. Notably, the default population size is comparatively small to allow for fast convergence. Restarts with increasing population size have been demonstrated [21] to be useful to improve the global search performance, and are nowadays included as an option in the standard algorithm.

In this research we used the CMA-ES package developed in R [10].

Table 3. The height of the RFU alleles peak (h instead of a) in each STR for each cow.

RFU	STR1	STR2	STR3	...	STRn
COW1_h	$h^{(1,1)x}$,$h^{(1,1)y}$	$h^{(1,2)x}$,$h^{(1,2)y}$	$h^{(1,3)x}$,$h^{(1,3)y}$	...	$h^{(1,n)x}$,$h^{(1,n)y}$
COW2_h	$h^{(2,1)x}$,$h^{(2,1)y}$	$h^{(2,2)x}$,$h^{(2,2)y}$	$h^{(2,3)x}$,$h^{(2,3)y}$	...	$h^{(2,n)x}$,$h^{(2,n)y}$
COW3_h	$h^{(3,1)x}$,$h^{(3,1)y}$	$h^{(3,2)x}$,$h^{(3,2)y}$	$h^{(3,3)x}$,$h^{(3,3)y}$	...	$h^{(3,n)x}$,$h^{(3,n)y}$
...	...	...	...	...	...
COWm_h	$h^{(m,1)x}$,$h^{(m,1)y}$	$h^{(m,2)x}$,$h^{(m,2)y}$	$h^{(m,3)x}$,$h^{(m,3)y}$	...	$h^{(m,n)x}$,$h^{(m,n)y}$
BM_h	$\underline{h^1_{BM}}$	$\underline{h^2_{BM}}$	$\underline{h^3_{BM}}$	...	$\underline{h^n_{BM}}$
CH_h	$\underline{h^1_{CH}}$	$\underline{h^2_{CH}}$	$\underline{h^3_{CH}}$	...	$\underline{h^n_{CH}}$

2.2.2. Computer-assisted molecular traceability pipeline

In this study we assume that, if a certain number of cows that produced the BM or CH does exist, then the BM or CH genetic STR profile should be a linear combination of the STR profiles of those cows. Under this postulate, the automated forgery detection we propose is composed of two steps: data normalization, and heuristic simulation.

The purpose of the data normalization step is to preprocess the RFU raw data (see Table 3) of a specific dairy product (CH or BM pool analysis) and the ones from the profiles of the cows belonging to the declared farm. This in turn makes them comparable and allows us to perform forgery detection. All RFU peak profiles are therefore normalized between [0,1] producing the normalized dataset reported in Table 4 where:

$$H^{(i,j)} = \left[\frac{h^{(i,j)x}}{max(h^{(i)x})}, \frac{h^{(i,j)y}}{max(h^{(i)y})} \right] \tag{1}$$

is the normalized pair values of alleles' RFU peaks for cow i and STR j;

$$\underline{H}_p^{(j)} = \frac{\underline{h}_p^{(j)}}{max(\underline{h}_p)} \tag{2}$$

is the normalized vector of alleles' RFU peaks for pool P (BM or CH) and STR j.

Table 4. Normalized cows and pool (BM and CH) STR-RFU peak tabular data.

Normalized	STR1	STR2	STR3	...	STRn
COW1_H	$H^{(1,1)x}$,$H^{(1,1)y}$	$H^{(1,2)x}$,$H^{(1,2)y}$	$H^{(1,3)x}$,$H^{(1,3)y}$	...	$H^{(1,n)x}$,$H^{(1,n)y}$
COW2_H	$H^{(2,1)x}$,$H^{(2,1)y}$	$H^{(2,2)x}$,$H^{(2,2)y}$	$H^{(2,3)x}$,$H^{(2,3)y}$	...	$H^{(2,n)x}$,$H^{(2,n)y}$
COW3_H	$H^{(3,1)x}$,$H^{(3,1)y}$	$H^{(3,2)x}$,$H^{(3,2)y}$	$H^{(3,3)x}$,$H^{(3,3)y}$	...	$H^{(3,n)x}$,$H^{(3,n)y}$
...	...	...	...	...	...
COWm_H	$H^{(m,1)x}$,$H^{(m,1)y}$	$H^{(m,2)x}$,$H^{(m,2)y}$	$H^{(m,3)x}$,$H^{(m,3)y}$	...	$H^{(m,n)x}$,$H^{(m,n)y}$
BM_H	$\underline{H^1_{BM}}$	$\underline{H^2_{BM}}$	$\underline{H^3_{BM}}$	...	$\underline{H^n_{BM}}$
CH_H	$\underline{H^1_{CH}}$	$\underline{H^2_{CH}}$	$\underline{H^3_{CH}}$	...	$\underline{H^n_{CH}}$

The proposed forgery detection heuristic works analyzing the normalized data reported in Table 4. Our technique assumes that the amount of milk from each cow used in the production of the analyzed dairy product is unknown. The goal of the heuristic is to find the best cows' weighted combination (W) in such a way that the sum of the weighted cows' STR profiles produces a pattern as similar as possible to those of the analyzed dairy product. As an output score, the proposed model returns the sum of the squared errors (SSE) of the differences between the alleles of the expected milk or cheese STR profile and the predicted one, multiplied by two penalty coefficient. The first penalty (P1) is the percentage of alleles that are included in the STR profile of the dairy product but that are not present in any STR cow profile. The second penalty (P2) is the percentage of alleles available in cow profiles but not detected in the genotyping process of the pool. In other words, P1 represents the possible introduction of a forgery, while P2 estimates the loss of alleles from the cows pattern

due, for example, to the ripening process or the sample collection procedure. The outline of the proposed method is shown in Figure 1.

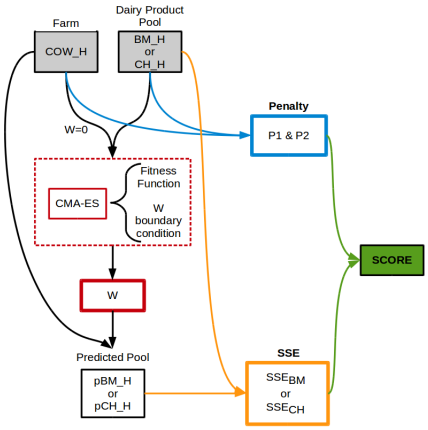

Figure 1. Global scheme of the Forgery Detection Model

The algorithm receives two main inputs:

- COW_H is the m×n matrix containing all normalized data for the cows composing the farm (Table 4). This table includes all data required to identify the target production farm for the dairy product under investigation;
- BM_H/CH_H is a vector reporting the normalized STR RFU peaks for the diary product under investigation (BM or CH) following the format reported in Table 4.

As a first step, the algorithm exploits the optimization capability of the CMA-ES to search for the best linear combination of the STR RFU peaks of the cows composing the farm (COW_H) able to generate the STR RFU profile of the diary product under investigation (BM_H or CH_H). This actually translates into the computation of a vector W of size m representing the computed contribution of each cow to the target diary product. Essentially the CMA-ES starts with an unknown weight vector equal to 0 (W=0). The CMA-ES then works over several generations until a stop condition is reached: max number of iterations or convergence. The best solution W identified by the CMA-ES is finally used to calculate the predicted profile for the target diary product as:

$$pP = W \times COW\_H \quad \text{where} \quad pP \in \{pBM\_H, pCH\_H\} \tag{3}$$

The computed profile (pP) and the original pool profile (BM_H or CH_H) can then be compared to calculate the sum of squared error (SSE_{BM} or SSE_{CH}) between the two profiles. This errors, corrected by the two penalty scores P1 and P2, can then be used to compute the final forgery score of the diary product with respect to the selected farm as:

$$\boxed{SCORE = SSE_P \cdot P1 \cdot P2} \quad SSE_P \in \{SSE\_BM, SSE\_CH\} \quad SSE\_BM = SSE(BM\_H, pBM\_H) \quad \text{and} \quad SSE\_CH = SSE(CH\_H, pCH\_H) \tag{4}$$

Since in case of frauds it may happen that a certain allele that appears in a specific STR of BM or CH does not appear in any STR allele of the cows, the RFU peak of that allele is taken into account in the SSE computation against a default value equal to 0. On the other hand, if the occurrence of a certain allele in a STR of a cow does not appear in the STR alleles vector of the pool, the routine automatically inserts a default value equal to 0 for that allele in the pool's STR vector. This last circumstance is possible when, during the genotyping process, or due to the ripening of the cheese, some allele are lost or not amplified enough.

The heuristic simulation is expected to return a score as close as possible to 0 in case of appropriate matching between the dairy products and the cows of a farm. Otherwise, in case of frauds, we expect that the automatic forgery detection returns a higher score value. In fact, in this case, there should be much more inconsistency in the match due to incoherent cows vs. dairy product STR patterns.

In order to perform its optimization, the CMA-ES requires the definition of a fitness function. Essentially, our goal is to minimize the SSE between the BM or CH genetic profile and the corresponding predicted one computed as a linear combination of the cows profiles. The SSE can therefore be exploited as an efficient fitness function for our goal. The temporary weight vector that is generated iteratively during the generation (g) is multiplied by the cows' profile to predict the temporary pool's pattern. The fitness function returning the SSE value is computed as follows:

$$\boxed{Fitness = SSE_P^{(g)}} \tag{5}$$

$$SSE_P^{(g)} \in \left\{SSE_{BM}^{(g)}, SSE_{CH}^{(g)}\right\}$$

$$SSE_{BM}^{(g)} = SSE(BM\_H, pBM\_H^{(g)})$$

$$SSE_{CH}^{(g)} = SSE(CH\_H, pCH\_H^{(g)})$$

$$pBM\_H^{(g)} \text{ or } pCH\_H^{(g)} = W^{(g)} \times COW\_H$$

$W^{(g)}$ is the temporary weight vector computed by CMA-ES at the generation g of the optimization process

One more important feature that was implemented in the software concerns W. Since in a farm, during the lactation period, each cow contributes with an unknown amount of milk w (that is essentially what the heuristic routine tries to estimate), we assume that every contribution cannot fall outside a predefined range that is:

$$lower\_boundary < w < upper\_boundary \tag{6}$$

$$lower\_boundary = \frac{0.5}{m} \quad \text{and} \quad upper\_boundary = max(\frac{3}{m}, 1)$$

The weight boundary condition is shown in Figure 2. It accounts for the fact that a cow cannot produce under/over a specific milk rate in relation to the number of the other milking cows (m). These constraints were chosen after analyzing several bulk milk batches and also after several discussions with the farm and veterinary staff. Basically it is supposed that each cows should produce more than a half and less of the triple of the mean quantity of the dairy product (i.e. $1/m$). Moreover, the upper boundary cannot exceed the value 1 since a cow must not produce all the dairy product by itself.

Anyway these constraints can be freely changed and they could be used to further refine the analysis in case of explicit information from producers concerning a particular dairy product.

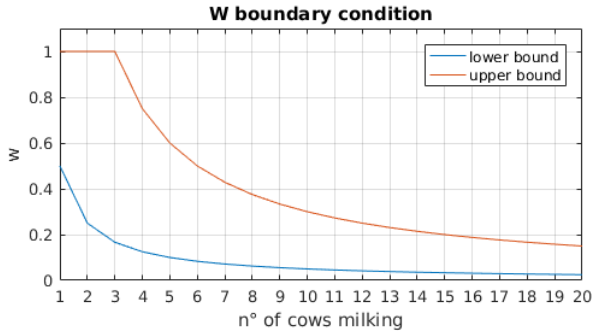

Figure 2. Boundary condition for w during the CMA-ES routine

2.2.3. Experimental Setup

To demonstrate the usability of the proposed approach we designed three experiments. The first one consists in analyzing the dairy product produced with 100% of milk of the same farm (i.e., COW_H, BM_H or

CH_H taken from the same farm). In the second experiment, instead, we analyzed a partial forgery in which a dairy product is produced from 50% randomly selected cows from a farm and 50% randomly selected cows from the other farm. Finally, in the third experiment, we analyzed a full forgery scenario in which we compared the diary product from a farm against the STR profile of the cows of the second farm.

For each farm and for each of the three forgery levels, every dairy product has been analyzed 24 times to highlight possible variations within the results. The whole experiment was executed in parallel on an eight-core machine Intel Xenon CPU E5-2680 @ 2.70GHz, 64 GB RAM, Ubuntu 14.04 LTS.

The STR Dataset previously described in section 2.1 is summarized in Table 5.

Table 5. Summary of the STR dataset used in the analysis.

Farm	No. Cows	No. Pool Samples
A	12	Bulk milk: 12 Derived Cheese: 12
B	14	Bulk milk: 11 Derived Cheese: 11

3. RESULT AND ANALYSIS

The main purpose of this work was to develop a new automatic methodology to highlight possible adulterations in dairy products thanks to a computational heuristic analysis. Using the method described in the previous sections, we obtained the results reported in Figure 3 and Figure 4.

Figure 3 reports the mean score values computed by the proposed heuristic over the 24 repetitions for the Bulk Milk analysis in Farm A and B. For each sampled pool, and for each month, the Figure shows the estimation of the three experimental setups described in section 2.2.3. with the changing forgery percentage. Figure 4 reflects the results of the cheese forgery simulation following the same criteria of Figure 3.

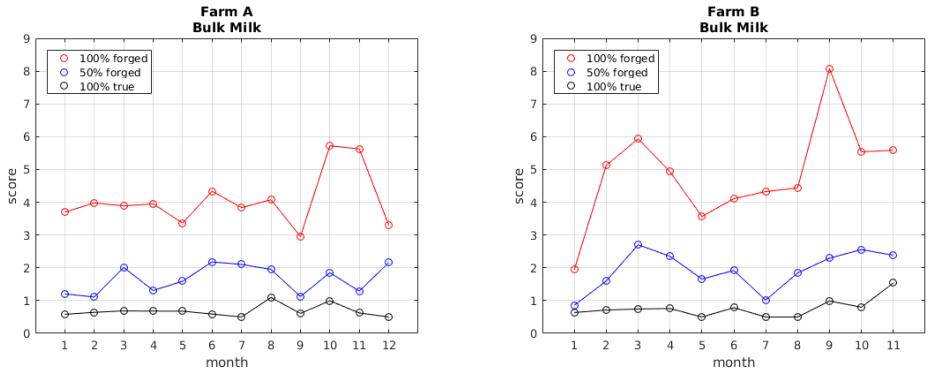

Figure 3. Results of the mean score values for Farm A (left side) and Farm B (right side) for the BULK MILK analysis for each available month. Black lines are related to 100% true cows setup analysis, the blue ones are related to 50% of adulterated milk origin, and the red ones are 100% forged milk origins.

Our forgery scores are overall very good according to the expected results: higher scores in case of adulteration and scores close to 0 otherwise. Moreover, it can be seen that partial forgery simulations are globally between 100% forged and 100% true examples. This behavior can be observed both in milk and cheese predictions. In the majority of the cases, the proposed automatic forgery detection reveals a considerably good accuracy with the exception of a few examples.

A summary of the aggregated results is given in Figure 5. These box plots represent the grouped results of Figure 3 and Figure 4, respectively. In general, the scores obtained for milk and derived cheese simulation indicate that it is possible to characterize our model with progressive cut-offs able to identify if forgery has occurred. As indicated in the figure the bulk milk boxes are noticeably well separated, while the cheese boxes show a less sharp separation in particular in the Farm B between the 50% forged and the 100% true group. The suggestion is that probably the STR profiles of the Farm A, that occur in the random selection

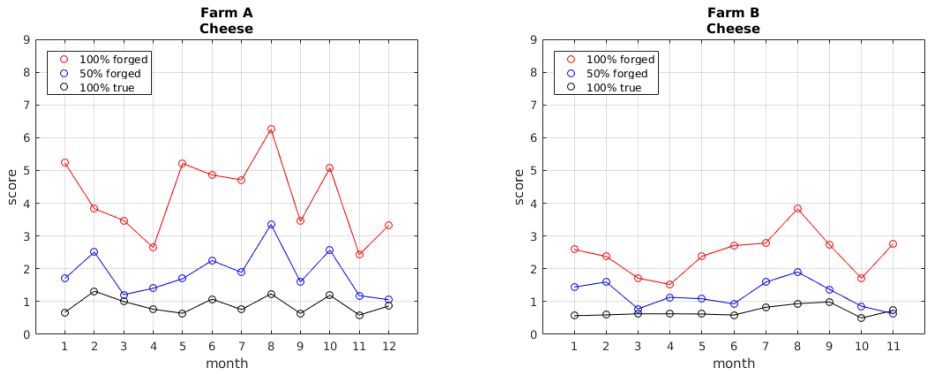

Figure 4. Results of the mean score values for the Farm A (left side) and the Farm B (right side) for the CHEESE analysis for each available month. Black lines are related to 100% true cows setup analysis, the blue ones are related to 50% of adulterated cows and the red ones are 100% forged cows.

for false cows, are too similar to the correct ones and only with a higher percentage of forgery the scores are extensively revealed.

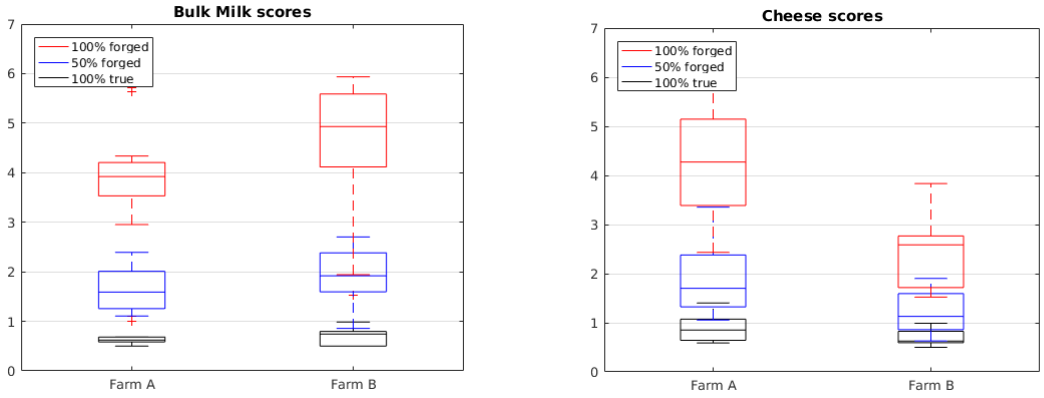

Figure 5. Box plots of grouped scores for the Farm A and B in the bulk milk and cheese analysis. Black box are related to 100% true cows setup analysis, the blue ones are related to 50% of adulterated cows and the red ones are 100% forged cows.

The overall results for the dairy product analysis is shown in Figure 6. Here the global scores are grouped together only to show the differences among the true simulation and the other two ratios of adulteration. Notice that Farm A and Farm B are merged, just like BM and CH.

The difference between the three groups (100% true, 50% forged and 100% forged) is statistically significant ($p<0.05$). This result also proves that it is possible to define a cut-off between distinctive levels of dairy product counterfeiting score (e.g., score=1 define adequately the limit for not forged product against half or complete falsified ones, score=2.5 is an opportune cut for sure complete falsification).

From the obtained results, it is evident that the automatic forgery detection model implemented and described in this paper is capable to identify the occurrence of irregular dairy product manufacturing and is also able to quantify the magnitude of the fraud. These results also suggest that this methodology may provide a useful strategy eligible to other food traceability context.

4. CONCLUSION

In this paper we proposed an innovative automatic forgery detection method based on a heuristic procedure. This system is able to measure the likelihood that a traditional dairy product is originated from a known farm, thus providing a measure of the level of potential counterfeiting. We investigated the use of Short

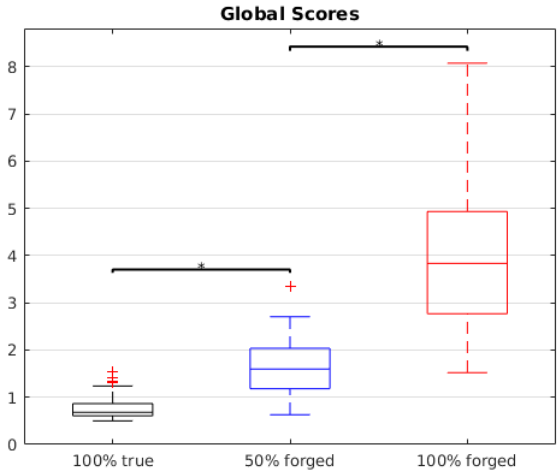

Figure 6. Global simulation scores. Both farms and dairy products are grouped. Black box is related to 100% true cows setup analysis, the blue one is related to 50% of adulterated cows and the red one is 100% forged cows. The * indicate significant difference between groups (p¡0.05).

Tandem Repeats associated to their relative fluorescence unit (RFU) to estimate the quantity of each individual that contributed in the final pool. We employed a Covariance Matrix Adaptation Evolution Strategy algorithm in order to predict the traceability between dairy products and the corresponding producer. Results obtained in several experiments provided excellent outcomes and encourage the research community to investigate further to employ this method to other foodstuff traceability issues.

ACKNOWLEDGEMENT

This work was supported by Italian Ministry of Health grant IZS PLV 01/14 RC.

REFERENCES

[1] H. Megens, et al., "Biodiversity of pig breeds from China and Europe estimated from pooled DNA samples: differences in microsatellite variation between two areas of domestication," *Genetics Selection Evolution*, vol. 40, no. 1, pp. 103-128, 2008.
[2] H. Schnack, et al., "Accurate determination of microsatellite allele frequencies in pooled DNA samples," *European Journal of Human Genetics*, vol. 12, no. 11, pp. 925-934, 2004.
[3] G. Skalski, et al., "Evaluation of DNA Pooling for the Estimation of Microsatellite Allele Frequencies: A Case Study Using Striped Bass (Morone saxatilis)," *Genetics*, vol. 173, no. 2 , pp. 863-875, 2006.
[4] C. Likhitha, P. Ninitha, V. Kanchana, "DNA Bar-coding: A Novel Approach for Identifying an Individual Using Extended Levenshtein Distance Algorithm and STR analysis," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 6, no. 3, pp.1133-1139, 2016.
[5] M. Widyanto, R. N. Hartono, N. Soedarsono, "A Novel Human STR Similarity Method using Cascade Statistical Fuzzy Rules with Tribal Information Inference," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 6, no. 6, pp. 3103-3111, 2016.
[6] A. Bagnato, et al., "Quantitative Trait Loci Affecting Milk Yield and Protein Percentage in a Three-Country Brown Swiss Population," *Journal of Dairy Science*, vol. 91, no. 2, pp. 767-783, 2008.
[7] E. Lipkin, et al., "Quantitative Trait Locus Mapping in Chickens by Selective DNA Pooling with Dinucleotide Microsatellite Markers by Using Purified DNA and Fresh or Frozen Red Blood Cells as Applied to Marker-Assisted Selection," *Poultry Science*, vol. 81, no. 3, pp. 283-292, 2002.
[8] J. Park, et al., "Determination of the Authenticity of Dairy Products on the Basis of Fatty Acids and Triacylglycerols Content using GC Analysis," *Korean Journal for Food Science of Animal Resources*, vol. 34, no. 3, pp. 316-324, 2014.
[9] M. Sardina, et al., "Application of microsatellite markers as potential tools for traceability of Girgentana goat breed dairy products," *Food Research International*, vol. 74, pp. 115-122, 2015.

[10] H. Trautmann, O. Mersmann, D. Arnu, "cmaes: Covariance Matrix Adapting Evolutionary Strategy," R package version 1.0-11, 2011.
[11] Team RC, "R: A language and environment for statistical computing," Vienna, Austria: R Foundation for Statistical Computing, 2014.
[12] R. Toonen, S. Hughes, "Increased throughput for fragment analysis on an ABI PRISM 377 automated sequencer using a membrane comb and STRand software," *Biotechniques*, vol. 31, no. 6, pp. 1320-1324, 2001.
[13] J. Goudet, "FSTAT a program to estimate and test gene diversities and fixation indices (version 2.9.3)," Available: http://www. unil. ch/izea/softwares/fstat.html, 2001.
[14] J. Felsenstein, "PHYLIP (Phylogeny Inference Package)," Available: http://evolution.genetics.washington.edu/phylip.html, 2005.
[15] N. Crawford, "smogd: software for the measurement of genetic diversity," *Molecular Ecology Resources*, vol. 10, no. 3, pp. 556-557, 2010.
[16] M. Hubisz, et al., "Inferring weak population structure with the assistance of sample group information," *Molecular Ecology Resources*, vol. 9, no. 5, pp. 1322-1332, 2009.
[17] N. Hansen, A. Ostermeier, A. Gawelczyk, "On the Adaptation of Arbitrary Normal Mutation Distributions in Evolution Strategies: The Generating Set Adaptation," *ICGA*, 1995, pp. 57-64.
[18] N. Hansen, A. Ostermeier, "Completely Derandomized Self-Adaptation in Evolution Strategies," *Evolutionary Computation*, vol. 9, no. 2, pp.159-195, 2001.
[19] N. Hansen, S. Mller, P. Koumoutsakos, "Reducing the Time Complexity of the Derandomized Evolution Strategy with Covariance Matrix Adaptation (CMA-ES)," *Evolutionary Computation*, vol. 11, no. 1, pp. 1-18, 2003.
[20] I. Ismail,A. Hanif Halim, "Comparative Study of Meta-heuristics Optimization Algorithm using Benchmark Function," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 7, no. 3, pp. 1643-1650, 2017.
[21] A. Auger, N. Hansen, "A restart CMA evolution strategy with increasing population size," *Evolutionary Computation, 2005. The 2005 IEEE Congress on. IEEE*, vol. 2, pp. 1769-1776, 2005.