Forensic Science International: Genetics Supplement Series 5 (2015) e159–e161

Contents lists available at ScienceDirect

Forensic Science International: Genetics Supplement Series

journal homepage: www.elsevier.com/locate/FSIGSS

Algorithm aligen: Simple method of encryption and matching with genetic profiles in a system of STR identification database

G.C. Iannacone*

Molecular Biology and Genetic Laboratory, Institute of Legal Medicine and Forensic Science, Lima, Peru

ARTICLE INFO

Article history:
Received 23 July 2015
Accepted 16 September 2015
Available online 24 September 2015

Keywords:
DNA Database
STR
Genetic Profile
Missing Persons
DNA identification
ALIGEN

ABSTRACT

During the DNA identification process of 15000 missing persons in Peru between 1980 and 2000, we observed many cases of random matches due to the population genetic structure (founder effect, low gene flow between communities and inbreeding). In this genetic context, since 2002, we have been developing an algorithm named ALIGEN with the aim of improving the match and identification. This algorithm performs a meiosis simulation in two DNA databases (relatives and missing persons). In each DNA database ALIGEN generated the haploid profiles (hap-file) for each genetic profile divided in five groups of four STR markers (match group) with a total of twenty STR markers. Simultaneously, we performs a kinship analysis using the model of allele Identity by descendent (IBD) using a threshold of 90% posterior probability in the case of fullsibs with the aim to obtain only the significative relationship and avoiding the random matches. To support the first analysis, the algorithm generated a genetic distances between two genetic profile using hap-file and this form the matrix of likeness that correspond to the genetic distance among all genetic profiles (relatives and missing persons). This matrix is used in MEGA with the aim to obtain a relationship tree. In this way we can confirm the matches, the random matches and evidence of unknown relationships. Other characteristic of the algorithm, it can able to ensure the DNA information privacy through the encryption of each genetic profile using a Hash encryption model with de hap-file generated and it is a criteria very important in the future of populations DNA databasing. Finally, ALIGEN have been validated in the last 13 years solving the identification of missing persons in many cases at national and international level. For this raison, we wish share this algorithm as an easy tool that it can be implementing for any laboratory using the formulas described in this article.

1. Introduction

In 2002 ALIGEN algorithm was initially designed with the purpose of reconstructing genealogies in wild and domestic animals [1]. In that year, it was used by the Institute of Legal Medicine of Perú to make possible the identification process of missing persons in an open massive case, in which we had the aim to process a total of 400 burn bones samples and about 186 family groups (a total of 458 relative). Later, this algorithm was used in the DNA identification in air crashes cases in Perú (2003–2007). Today, it is used in the process of DNA identification of missing persons in Perú between 1980 and 2000 (around 15000 missing persons to be identified). In the other hand, at international level, we used the ALIGEN algorithm, with the aim to help the Government of Honduras in the process of identifying of burn bodies in a prision of Tegucigalpa in 2012.

2. Materials and methods

The algorithm ALIGEN was developed using visual basic in Excel. The basis of this system is generate all possible haploide profiles (hap-file) in one STR genetic profile. Simultaneously, ALIGEN calculate the kinship probability using the IBDs methodology [2] and the prior probability in relation to the number of missing persons in the DNA database [3]. With the aim to improve the quality of matches and avoid the random match, ALIGEN include a matrix of likeness of hap-file. Then, the method considers the flow criteria:

a. Indexed all the matches with threshold of a 90% of posterior probability with the criteria of Fullsibs for each duo of relative and missing person in the DNA database. Also, we can know at

* Corresponding author.
E-mail address: ggiannacone@yahoo.com (G.C. Iannacone).

http://dx.doi.org/10.1016/j.fsigss.2015.09.064
1875-1768/

the same time, if there are complete match or fail at least one marker (case of mutation).
b. Contrasts the positive matches with the Matrix of likeness of hap-file tree construct with MEGA V4.0.2. This with the aim to confirm the marches, random match as in the case of Sub-structure and evidence relationships unknown.

3. Results and discussion

3.1. Considerations in the developing of the ALIGEN algorithm

a. We assume that the markers used are no linked, diploid and polymorphic.
b. Each meiotic combination is called haploid profile (Hap-File). This Hap-File follows a multiplicative principle. Where in a (r) STR markers there are two state (homozygous or heterozygous), then corresponding to each STR marker state: r1, r2, ...rn. Therefore, the number of combinations of Hap-File would be equal to 2^r.
c. The alleles values are expressed as X1, X2,..,Xn, which correspond to each allele in the marker 1, 2, 3, ... n, and it have an specific position predetermine by the user as N1, N2, ...Nn which correspond to the value of 1, 2, 3 . . . n.. Then, the specific weight of each hap-file is represented as: $\sum_{n=1}^{r} N_n.X_n^2$..
For example if we have the Hap-File of 4 alleles in the follow fixed position: Marker 1 allele 12, Marker 2 the allele 30.1, Marker 3 the allele 9.3 and Marker 4 the allele 16; the value of hap-file will be $1x(12)^2 + 2x(30.1)^2 + 3x(9.3)^2 + 4x(16)^2 = 3239.49$
d. The specific weight of each Hap-File in an individual are additive.
e. We can form groups of Hap-Files (match group) for the STR markers considered that form the p match groups. For example, if we build a DNA database with 20 STR markers, we can build five match groups ($p = 5$) of 4 STR markers in each p match group.

3.2. Implementation of the ALIGEN algorithm in Excel to generated the hap-file

a. One sheet is to generating the hap-File corresponding to query genetic profile of missing person. With a total of 16Hap-File for 1 match group of 4 STR markers. It is the same for the relative in other sheet. Then, the total groups generated are 5 match groups with a total of 20 STR markers.
b. For comparison of Hap-File there is one sheet, where it have the results of the comparisons of Hap-File generated in the point "a" between the duo relative and missing person. Then, in each match group, we will have a total of 256 comparisons between the two STR genetic profiles. The positive match have values greater than zero in the five match groups. These results can be indexed to show only those genetic profiles that have a positive match. Also, if the matching fails in a one match group, we would be facing a probable mutation, where we suppose a priori at least in these match groups fail at least in one STR marker. Similar to a full match, we can indexed to find those genetic profiles that fail in a match group.

3.3. Strategy of missing person identification in massive cases

With the aim to improve the quality of the match, ALIGEN have two levels of analysis. One level include the probability of two profiles share alleles identity by descendent (IBD) and the second level include a matrix of likeness of hap-file.

In the case of IBD, we consider indexed the results using the criteria of relatedness of Fullsib (K2 = 0.25, K1 = 0.5, K0 = 0.25) or Parent–Child (K2 = 0, K1 = 1, K0 = 0) to avoid the probability of random match. In this level, it is important choose the adequate threshold. It is recommendable indexed by Fullsib results with a threshold of posterior probability of 90%. This permit show only the significative relatedness and avoid the cases where it have a positive match or fail at least in one marker due to it share alleles frequent at the population level (random match). At the same time there are a second kinship calculation as for example in this case a Father–Child or Mather–Child relationship.

To support the first analysis the second level in ALIGEN include a matrix of likeness of hap-file. Each hap-file are weight using the expression $Y = \sum_{n=1}^{r} N_n.X_n^2$ as we mention above. In the case of a match, all the hap-file values in one match group are weight in one value as $\sum_{n=1}^{q} Y_n$,where Y_n correspond (n) values of hap-file and (q) it is equal to (rn) hap-files generated. For the p match groups it is equal to $\prod_{n}^{p}\sum_{n}^{q} Y_n$. Then, the total value in the five match group are converter in a number less than 1 with the expression $(Ln((\prod_{n}^{p}\sum_{n}^{q} Y_n) + 1))^{-1}$ where a big value it is converted to a value near to 0, thus, it correspond at the minor distance between two genetic profiles and inverse in the case of small values. With the aim to have a graphic relationships in the DNA database (relative and missing person), the matrix it is analyzed with MEGA software to obtain a tree of relatedness. In this way we can confirm the relationships founded with the IBD and analyze if there are random match and/or unknowns relationships in the case.

For more information enter in the link below:
https://www.dropbox.com/s/4hkig1wcyk5yb35/ALIGEN%20V4.1%20beta%20Ok.xlsm?dl=0

3.4. Getting the genetic identification value (GIV)

The ALIGEN algorithm assigned a unique value called Genetic Identification Value (GIV). It correspond a number related to allelic variability in the STR genetic profile. This value is obtained by $\prod_{n=1,3..}^{p} (G_n + G_{n+1}) \times 10^{-z}$, where we assuming that each group corresponds to (p) match groups with a given set of markers, where n is the nth generated match groups, Gn and Gn +1 correspond to each value of each match group formed for the 16Hap-File in an one match group of 4 STR markers and (z) it is for adjust the range of digits required. According to the five match groups (20 markers), we have $p = 5$ and an we need a GIV of 14 to 16 digits then the expression for the calculation of GIV for each profile will be $(G_1 + G_2) \times (G_3 + G_4) \times (G_5 + 0) \times 10^{-1}$. The decimal adjust value of GIV is your top number from 0.5 to 0.99 and the lower value between 0.01 and 0.49.

This GIV meets the following properties of the Hash model:

a. Whatever the length of the base A (genetic profile) text, the length of the resulting Hash B (GIV) will always be the same.
b. For each A, the function generates a single output B

c. Given a text base, is easy and fast (for a computer) calculate the value
d. It is impossible to reconstruct the text based (genetic profile) from abstract the value (GIV)

4. Conclusion

This is a tool that contributes to the human identification and database algorithms that exist in the genetic forensic community. Also, this can be easily designed in each laboratory to solve problems of identification in masive cases and it is a basis to generate their own search and encryption system in cases of implementation of DNA database program applied to human identification.

Conflict of interest

None

Role of funding

The research was conducted with funding of the Public Ministry and author.

Acknowledgements

The research was made possible thanks to the support of the Attorney General of Peru Lima–Peru.

References

[1] G.C. Iannacone, Use of Microsatellites markers to determine the paternity of Alpaca Huacaya Race. Universidad Nacional Agraria la Molina, Magister Scientiae These 2005.
[2] B. Weir, A. Anderson, A. Hepler, Genetic relatedness analysis modern data and new challenges, Nat. Genet. 7 (2006) 771–780.
[3] W. Goodwin, C. Peel, Theorical value of the recommended expanded European standard set of STR loci for the identification of human remains, Med. Sci. Law (2012) 162–168.