Searching a DNA Data Bank Using Complex Mixtures; a Retrospective Study.

Lavergne Léo, Mailly France, Noël Josée and Jolicoeur Christine
Laboratoire de sciences judiciaires et de médecine légale, Montréal, Québec, Canada

Abstract

Since the establishment of the National DNA Data Bank of Canada in 2000, and the implementation of the CODIS software in our laboratory, we have developed a working routine with complex DNA mixtures in order to find valuable candidate matches at the local and national levels. Mixture profiles (maximum two contributors) with up to 6 mixed loci (out of 13 P+/COfiler®) are uploaded to the National level, while mixtures with 7 or more mixed loci are kept at the local level. In order to better keep track of the match results for these complex mixtures, the samples are coded according to the number of mixed loci, and a 'px' extension identifies those derived from even more complex mixtures. We have examined all the match results (774 candidates) obtained to date (May 2008) with 2 400 mixed profiles searched against more than 130 000 convicted offender profiles. We compared the number of 'no match', 'conviction match' and 'offender hit' dispositions for different mixture categories and single-source profiles. Overall, the proportions obtained for the various match dispositions were similar for mixture and single-source profiles. With this data, we understand better what one can expect from searching such mixtures. It is clear that valuable matches can be obtained using complex mixtures provided that key elements are integrated to the procedure. Good mixture interpretation guidelines should be coupled with databanking strategies designed to maximize the probability of a valuable match while minimizing the risk of adventitious hits. Also, downstream review of matches using the original electrophoregram data should be carried out by experienced analysts. With such a setup, we have been able to obtain many matches that would not have been identified otherwise. Further studies to analyse peak balance of matching/non-matching alleles within individual loci and to search complex mixtures against larger databases are under way.

Introduction

Mixtures were less of a concern in the early days of DNA profiling (1). With the implementation of STR analysis and technologies of increasing sensitivity, a large number of mixtures with different degrees of complexity can be obtained from a variety of sample types. Interpretation of DNA mixtures (deconvolution, extraction) is a major concern in the scientific literature and important steps have been taken in using more sophisticated computer analysis for mixture interpretation and calculation of their probative value (2-5). On the other hand, little has been said on the use of DNA mixtures in DNA data banks (6-7). Our organization (LSJML, under the jurisdiction of the Québec government, Ministry of Public Safety) has been working with DNA mixtures for databanking for 8 years.

Our laboratory receives all cases from the Québec province (population of about 7 millions) and handles more than 3 500 cases a year. By adoption of a federal bill in 2000, the CODIS-based

National DNA database of Canada (NDDB) was implemented with two indexes: a Convicted Offender Index (COI) (administered by the Royal Canadian Mounted Police (RCMP), and a Crime Scene Index (CSI). Local labs (Quebec, Ontario, other Canadian provinces) populate the Crime Scene Index with single source profiles (Forensic Unknown) and mixtures (Forensic Mixture). The CSI is searched against itself and against the COI. In 2001, our laboratory moved from the 9 ProfilerPlus® loci to the 13 CODIS core loci with the addition of the Cofiler® system. To date, our laboratory has uploaded nearly 11 500 profiles in the Forensic Unknown index, and 2 400 in the Forensic Mixture index, the vast majority analyzed using 13 loci.

Here we present data obtained by comparing the Québec Forensic mixture profiles to the COI of the Canadian National DNA data bank and also data from a subset of those mixtures compared to the Florida COI.

Approach

Since its implementation, the DNA data bank has been used in our laboratory as a search tool and an aid to investigation. When relevant to the case, mixtures are interpreted and uploaded to CODIS. We aimed to develop databanking strategies for an optimal use of CODIS, maximizing the probability of finding the offender profile if present in the data bank, while minimizing the number of candidate matches that would end up with a no match disposition after time and energy had been spent in their review.

Quebec's DNA profiles uploaded at the National level in the Forensic Mixture index must not have more than 6 mixed loci out of the 13 CODIS loci, with no more than 4 mixed loci from the ProfilerPlus® system. We define a mixed loci as a loci with more than 2 alleles. In our laboratory, mixtures that can be entirely deconvoluted to no more than 2 alleles on all loci using our interpretation guidelines, are uploaded to the Forensic Unknown index. For complex, two-contributor mixtures (no well-defined major profile, no known contributor) that can be only partly deconvoluted, or when a third party is present, flexible databanking guidelines are used for uploading to the Forensic Mixture index. Mixtures of higher complexity are kept at the local level. This study focuses on mixtures in the Forensic Mixture index.

Mixture management

- Coding

An essential feature of the mixture management system in our laboratory is a specific coding system that we implemented for complex mixtures (≥4 mixed loci). Such mixed profiles are coded according to their degree of complexity using an alphanumeric system (representing different parameters) included as a suffix in the CODIS ID number. For example, in 08M1234LL2PX-7/6, PX indicates that alleles were left out from the original results for databanking, while 7 represents the number of loci with two alleles or less and 6 the number of loci with three alleles or more. The coding system allows easy

tracking down of the mixtures uploaded to the data bank, the candidate matches generated, as well as their match dispositions, according to their degree of complexity.

- Strategies for databanking

Databanking strategies begin upstream with a thorough evaluation of the quality of the mixture. In our hands, peak heights above 300 - 400 rfus are more reliable. The number of contributors and their respective proportions have to be estimated (8-9). A decision is made regarding which and how many significant contributors from the mixture will be included in the upload, based on relevant casefile information, relative proportion of contributors and peak heights. Two-contributor good quality mixtures are best for databanking; currently, mixtures with more than three contributors are not used, and the use of three-contributor mixtures is limited. For instance, mixtures with one main contributor and traces of two others, or two main contributors and traces of a third, may be used for databanking provided that peak heights and general quality of the profile are good.

Subsequently, interpretation protocols slightly more relaxed than our standard interpretation guidelines are used. Flexible databanking strategies for each locus, based on peak heights and relevant casefile information, may include:

- leaving out minor alleles and choosing a smaller set, or one specific allele (must limit the number of one-allele loci);
- using the CODIS obligatory allele to maximize the discriminatory power of a set of alleles;
- leaving out high molecular weight loci (generally no more than 2) when the probability of drop-outs is obvious.

Match disposition criteria

Downstream, each moderate stringency candidate match obtained is reviewed by a CODIS administrator with the electrophoregram in hand taking also consideration of other case work information (other DNA profiles). Using a set of flexible match disposition criteria, we try to exclude the candidate:

- Candidate (offender) should be present in the target (mixture). Drop-outs are unacceptable for the main contributor.
    - For the lower proportion contributor, drop-outs may occur (generally no more than 2), but must be consistent with peak heights and molecular weights (ex.: D18, D7, CSF1PO). In general, they should be visible under the threshold, and when so, they must be at expected positions.
- The putative presence of the candidate in the mixture must be consistent with the previously estimated number of contributors.
- The putative contribution of the candidate to the mixture must be consistent with peak heights and the estimated proportion of contributors across all or most loci.

Each moderate stringency candidate match is unique and reviewed on its own value. Candidate matches for which no inconsistencies could be found, or where missing criteria on one or two loci can be easily explained, receive an offender hit (or forensic hit) disposition. Others receive a no match disposition.

Searching a Convicted Offender Index (COI) with complex mixtures

As of June 2008, Quebec had populated the Forensic mixture index with 2 392 mixtures, one third having 4, 5 or 6 mixed loci (out of 13). In the present study, we looked at the behaviour of these mixtures when run against the COI. Using our codification system, we traced them back according to their degree of complexity and compared their match disposition outcome. We also compared their match disposition outcome to single-source profiles.

Table 1 presents the results obtained for the last 8 years for low complexity mixtures (≤ 3 mixed loci), moderate complexity mixtures (4 to 6 mixed loci), and profiles from the Forensic Unknown index. For each category are shown the total number of candidate matches obtained, and the number of candidates that received one of three types of match disposition: conviction match[1], no match or offender hit.

The data show that mixtures with 4 to 6 mixed loci (out of 13) do not produce more candidate matches in proportion than mixtures of lesser complexity or profiles from the Forensic Unknown index. Approximately a fifth of moderate complexity mixtures (178 / 885) produced candidate matches, while a third of low complexity mixtures (454 / 1 507), and a little less than half of Forensic Unknown profiles (4 897 / 11 455), produced candidate matches.

Moreover, the match disposition outcome of these candidates is comparable between the three categories. The proportion of conviction match dispositions is nearly identical between candidates from low complexity mixtures, moderate complexity mixtures and Forensic Unknown profiles (~ 25% of candidate matches). For this match disposition, the same identification is obtained by two independent processes, namely the judicial system and the mixture search in the data bank, providing evidence of the efficiency of our banking strategies and match criteria. Similarly, less than 15% of reviewed candidate matches received a no match disposition for all three categories. Finally, searching these profiles against the COI provided valid investigative leads in similar proportions for all three categories, with ~ 55% - 65% of candidate matches receiving an offender hit disposition across categories. Clearly, adequate mixture databanking strategies can maximise the use of a data bank as an investigative tool, while keeping manageable the proportion of reviewed candidate matches that end up with a no match disposition.

[1] A conviction match disposition is given when the CSI profile and the national COI profile relate to the same offence.

It is of note that for the 117 offender hits obtained from mixtures with 4 to 6 mixed loci, nearly half (57) did not produce any candidate match with the local Crime Scene Index (CSI). The only candidate match obtained from each of these 57 mixtures was with the Convicted Offender Index at the National level (data not shown). Therefore, had the moderate complexity mixtures not been uploaded to the National level but been kept locally, the only investigative lead provided by the data bank would have been lost for 57 samples.

Exploring the limits of searching a data bank with complex mixtures

We were interested in exploring further the use of mixtures in data banks by testing mixtures of higher complexity (7 – 13 mixed loci out of 13), or by searching moderate complexity mixtures against larger size data banks. In this preliminary study, three small-scale searches were designed.

First, we used 38 moderate complexity mixtures that each had produced one candidate match with the COI. We added back all minor alleles from the original results left out for databanking, resulting in more complex mixtures with 7 to 13 mixed loci and having up to 5 alleles/loci. This subset of profiles was sent to the National level for a new search against the COI (125 000 profiles by June 2007). In addition to the 38 candidate matches observed previously, only 2 mixtures each matched one additional offender (both of these candidate matches receiving no match disposition upon electrophoregram review). No mixture produced numerous candidate matches.

Second, we asked the CODIS administrator at the National level to search 20 two-contributor high complexity mixtures (7 – 12 mixed loci), against the COI (130 000 profiles by June 2008). In this search, 10 mixtures returned 12 candidates (some already known), of which 4 received a no match disposition upon review, while 10 mixtures remained silent. These preliminary results suggest that there is no drastic effect to expect when searching high-complexity mixtures with several alleles on several loci against a mid size data bank.

Third, we planned to search moderate complexity mixtures against a larger size data bank, and we turned to the Florida Department of Law Enforcement laboratory in Tallahassee, asking Dr. David Coffman for his collaboration in this study. By June of 2008, the Florida COI was populated with about 480 000 offender profiles. We sent the Tallahassee laboratory the 38 mixtures mentioned above, in their moderate complexity form (as originally submitted to the NDDB) (3 to 6 mixed loci), and asked the CODIS administrator to search them against the Florida COI. When tolerating no mismatch, only one of the 38 moderate complexity mixtures produced 2 candidate matches, the 37 other mixtures remaining silent. However, when tolerating one mismatch (no match at one locus), the results obtained were very different. In this case, two thirds (25) of the mixtures remained silent, while 13 mixtures returned a total of 95 offenders with one mixed profile generating 27 hits.

Taken together, these results strongly suggest that the number of loci is a critical factor in the discriminatory power of a mixture when searching against a data bank, and that it is, actually, more important than the number of alleles in each mixed locus

Discussion and conclusion

The Canadian NDDB currently holds more than 130 000 profiles in its COI and those are continuously compared to 2400 mixed profiles from the Québec local data bank. Since 2000, we have obtained close to 400 offender hits against mixture samples, with almost a third involving profiles having 4 to 6 mixed loci (moderate complexity) (table 1). The review of candidate matches carried out in June 2008 has shown similar match disposition proportions for single-source profiles, low complexity mixtures (3 or fewer mixed loci), and moderate complexity mixtures. Most of the mixed profiles (>60%) uploaded to the NDDB have not generated a single candidate match over the years (data not shown). Thus, complex mixed profiles can be used efficiently to search convicted offender databases containing well over one hundred thousand profiles. However, we believe that several key conditions must be met for this approach to be successful and worthwhile (or cost-effective).

First and foremost, a well-defined framework for profile and match management is required and is dependant on staff dedicated to these tasks. Casework analysts, with the support of CODIS administrators, must carefully evaluate complex profiles using a conservative approach for allele selection prior to databanking. Then, once a candidate match involving a mixed profile is returned, a thorough review must be carried out by a CODIS administrator that is himself a casework analyst with considerable DNA analysis experience. The review must take into account the original electrophoretic data as well as all other relevant case information (such as other genetic profiles observed in the case). Moreover, any accepted match should be submitted to the case analyst for final approval. Thus, this framework requires a significant time investment and intensive training of all analysts involved. In addition, we have found that the presence of specific office clerks devoted to casefile retrieval and communications with investigators is essential for the efficiency of the system.

The vast majority of matches in our study that were given a conviction match or offender hit disposition (acceptable matches/positive disposition) involved crime scene mixtures with data on all 13 CODIS core loci. It has been our experience that the comparison of single-source profiles having only Profiler+® data with 13-loci moderate complexity mixtures generate mostly no match dispositions (data not shown). Together with the preliminary data from the Florida database, single-mismatch search that returned multiple candidates from several mixed profiles, this suggests that having data on all 13 loci may be highly advisable when using mixed profiles of moderate to high complexity. Moreover, the number of alleles per locus appears to be less critical. This latter parameter is otherwise limited by the fact that a maximum of two significant contributors is generally allowed for databanking.

In conclusion, providing that the necessary structure is in place and that technical requirements concerning the number of loci and search parameters are properly considered, we believe that the

time and energy invested in the management of complex mixtures for databanking purposes is well-worth the effort.

Acknowledgments

We would like to thank the staff of the Canadian National DNA Databank in Ottawa, in particular Sylvain A. Lalonde, National CODIS manager, for their support and for handling all the results obtained with our mixed profiles at the National level. We also wish to thank for their kind collaboration Dr David Coffman, Chief of Forensic Services and Chris Carney, DNA database supervisor at FDLE in Tallahassee.

References

1- Evett IW, Buffery C, Wilcott G, Stoney D. A guide to interpreting single locus profiles of DNA mixtures in forensic cases. J Forensic Sci Soc 1991; 31: 41-47
2- Bill M, Gill P, Curran J, Clayton T, Pinchin R, Healy M, Buckleton J. Pendulum- a guideline-based approach to the interpretation of STR mixtures. For Sci Int 2005; 148: 181-189
3- Wang T, Xue N, Birdwell J D. Least-Square Deconvolution : A framework for interpreting short tandem repeat mixtures. J For Sci 2006; 51-6: 1284-1297
4- Cowell RG, Lauritzen SL, Mortera J. Identification and separtation of DNA mixtures using peak area information. For Sci Int 2007; 166: 28-34
5- Gill P, Brenner CH, Buckleton JS, Carracedo A, Krawczak M, Mayr WR, Morling N, Prinz M, Schneider PM, Weir BS. DNA commission of the International Society of Forensic Genetics: recommendations on the interpretation of mixtures. For Sci Int 2006; 160: 90-101
6- Torres Y, Flores I, Prieto V, Lopez-Soto M, Farfan MJ, Carracedo A, Sanz P. DNA mixtures in forensic casework: a 4 year retrospective study. For Sci Int 2003; 134:180-186
7- Voegeli P, Haas C, Kratzer A, Bär W. Evaluation of the 4-year period of the Swiss DNA database. International Congress Series 2006; 1288: 731-733
8- Paoletti DR, Doom TE, Krane CM, Raymer ML, Krane DE. Empirical analysis of the STR profiles resulting from conceptual mixtures. J For Sci 2005; 50: 1361-1366
9- Buckleton JS, Curran JM, Gill P. Towards understanding the effects of uncertainty in the number of contributors to DNA stains. For Sci Int Gen 2207; 1: 20-28

Table 1

Profile complexity	≤3 Mixed loci n=(1 507)	≥4 Mixed loci n=(885)	Forensic Unknown n=(11 455)
Candidate Matches	454	178	4 897
Conviction Matches	110 (24.2 %)	46 (25.9%)	1 245 (25%)
No Match	56 (12.3%)	10 (5%)	567 (11.6%)
Offender Hits	260 (57.2%)	117 (65.7%)	2 667 (54.5%)

Table 1: Comparison of low complexity (≤3 mixed loci) and moderate complexity (≥4 mixed loci) mixtures with single-source profiles (Forensic Unknown Index) according to three match dispositions. Similar proportions are observed across profile categories. The proportion of candidate matches obtained from these three groups (relative to the number of profiles) is respectively 30% (454/1507), 20% (178/885) and 42% (4897/11455).