International Congress Series 1261 (2004) 42–44

www.ics-elsevier.com

Large-scale comparative genotyping and kinship analysis: evolution in its use for human identification in mass fatality incidents and missing persons databasing

B. Leclair*

Myriad Genetic Laboratories Incorporated, Product Development, 320 Wakara Way,
Salt Lake City, UT, 84108, USA

Abstract. The discovery of mass graves and the advent of the World Trade Center tragedy have underscored the necessity to better harness the information content of large DNA typing data sets with bioinformatic tools. This presentation will review the evolution of a bioinformatic tool originally created to assist with the identification of the victims of the Swissair Air Disaster in 1998. This tool was re-built in 2001 to handle the much more demanding complexities of the WTC tragedy. This second generation included the capability to process very large STR and SNP data sets, to collapse remains data, to build composite profiles from overlapping partial remains profiles, to search for potential kinship associations on a purely genetic basis, to verify the consistency of complete family pedigrees, and to perform likelihood ratio calculations. The next generation of this tool is under construction as a stand-alone application under a new database architecture. It will provide functionalities to meet the needs of very large mass fatality incidents (MFIs), and of missing persons databasing initiatives. The considerable benefits of integrated, systematic "data mining" approaches will be discussed. © 2003 Elsevier B.V. All rights reserved.

Keywords: Mass fatality incidents; Missing persons databasing; Bioinformatics; Software; STR

1. Introduction

In this last decade, as DNA typing evolved and made it possible to process large numbers of samples, the successful use of the technology in the processing of remains recovered from mass fatality incidents (MFIs) has established the technology as a pivotal human identification tool for this type of event. DNA typing is unique in its ability to derive identity information from any type of tissue, which often proves crucial in incidents where body fragmentation is severe. DNA typing has also proven robust in generating genotyping results from remains exposed to severe environmental insult/decay. Despite these undeniable

* Tel.: +1-801-883-3814; fax: +1-801-584-1190.
E-mail address: bleclair@myriad.com (B. Leclair).

doi:10.1016/S0531-5131(03)01661-3

technical advances, large-scale massive fragmentation/decay MFIs still face significant complexities on the DNA data analysis front, as this type of field exercise rarely benefits from the availability of uncompromised remains, personal effects of reliable source attribution for all the victims, of reference samples from mother + father and/or offspring + spouse for all the victims. The necessity to work with very large but yet incomplete data sets renders the task of making accurate kinship associations more difficult. In that context, large-scale comparative genotyping is unavoidable. This manuscript reviews different iterations of a bioinformatic tool created to carry out such a function in large-scale MFIs.

2. The Swissair Flight 111 air disaster

The event took place in 1998, 229 passengers and crew perished, the crash dispersed thousands of severely fragmented human remains. Successful PCR amplification produced genotypes for 97% of the recovered remains [1]. Perfect matches with the genotypes derived from trace biological evidence recovered from personal effects purported to have belonged to the victims was immediately considered. However, as some of the most useful personal effects of the victims had been lost at sea in luggage, this route, although useful, was not perceived as being sufficient. Kinship analysis was considered, but both biological parents and/or offspring and spouse were available for only 26% of all victims. A total of 43 families including both parents with all or nearly all of their children were among the victims and represented another 26% of the victims. For the latter group, the necessity to reconstruct family pedigrees from within the questioned genotype data set was anticipated to be a demanding task. The anticipated encounter of core repeat slip mutations was also perceived as a complicating factor. A pair-wise genotype comparison scheme where a victim genotype is queried against the entire data set comprised of both next-of-kin references and other victims appeared to be a screening/sorting/scoring solution that could be rapidly implemented in an automated fashion for the rapid identification of significant biological relationships. On the basis of Mendelian inheritance rules, suspected parent/offspring relationships could be inferred by computing the number of loci at which at least one allele is shared. The scheme was automated with VBA in Excel. It permitted for all possible (i.e. 180,000) pair-wise genotype comparisons to be examined. As a closed population of victims was reached for this tragedy, the approach also permitted for the systematic exclusion of fortuitous kinship associations. Despite the design simplicity of the scheme, all inferred identification leads for the 218 victims for whom reference samples were available were supported by frequency and likelihood calculations carried out separately. This outcome confirmed the value of this screening approach as a first step in a victim identification strategy.

3. The World Trade Center tragedy

A very different situation was encountered with the WTC tragedy where nearly 3000 victims had perished, approximately 20,000 human remains were recovered, reference samples from 6000 relatives and 5000 personal effects were collected. A quick assessment with simulated data as to the suitability for the WTC situation of the approach used for the Swissair incident proved positive. However, the processing algorithms needed to be completely redesigned to reduce execution time by 150-fold for the 33×10^6 pair-wise genotype comparisons anticipated for the WTC data. This entirely new platform (MDKAP) included a variety of new functionalities, including the capability to: (1) collapse the 20,000 partial/complete genotypes derived from the remains to a reduced number of consensus genotypes; (2) assemble composite "virtual" genotypes from overlapping partial genotypes. A capability to search for a productive parentage trio was built in as well. However, evidence surfaced several months into the operation that for an unknown number of next-of-kin samples, the reported biological relationship to the victim could prove incorrect. In order to compensate for this deficiency, a triangulation approach was abandoned for a systematic testing, for each queried victim's genotype, of all potential trios that could be assembled with combinatorial pairs of relatives. This systematic approach, albeit computationally intensive, provided a matching process insensitive to potential errors in a sample's accessory information. The calculation of likelihood ratios for pair-wise comparisons was introduced, and was found to produce a very similar ranking of score results. The WTC victims' data set not being a closed population, fortuitous kinship associations could not be systematically eliminated. To alleviate this limitation, a feature was built in to provide the data reviewer with: (1) the score of a matching individual; (2) the score for each other member of the matching individual's family. Entries linked to fortuitous kinship associations systematically showed discrepant scores against the victim among their family relatives, and were easily eliminated from consideration by the software. The entire victim's genotype data set was queried in batch mode against the next-of-kin data set, and the results were stored. In order to facilitate and expedite data review, a prioritization scheme would then draft a list of the victim's genotype in decreasing rank for probability of yielding an identification. The batch process ran unattended and could be run as often as new data releases were produced. Inferred identifications were confirmed with additional pedigree analysis tools.

4. The future

The bioinformatic tools described above showcased the value of systematic "data mining" approaches that allow for optimal and efficient extraction of the biological information content of an available genetic data set. The next version of this tool is under construction and will address the processing needs of MFIs involving much larger numbers of victims, and of missing persons databasing initiatives. Built on a new database architecture, it allows parallel processing to keep execution time constant, accommodates over 10^6 samples, and will include a complex pedigree analysis package.

Acknowledgements

The author gratefully acknowledges George Carmody for the use of the KinTest program for likelihood ratio calculations.

Reference

[1] Leclair, B., Frégeau, C.J., Bowen, K.L., Fourney, R.M. Enhanced kinship analysis and STR-based DNA typing for human identification in mass fatality incidents: the Swissair Flight 111 disaster, submitted for publication, 2003.