

# RESEARCH ON THE DEEPENING APPLICATION OF FORENSIC DNA DATABASE BASED ON THE BIG DATA

Li Sheng, Cui Jie, Institute of Forensic Science, Ministry of Public Security

With the rapid development of artificial intelligence, mobile Internet and Internet of things, data resources have become a new asset, and the value of data is fully embodied in various fields. Human society has gradually entered the era of big data. Big data can enable people to grasp unprecedented comprehensive information. Effective analyses of these data can facilitate more accurate uncovering of the law of the development of things.

Over the last decade, China Forensic DNA database has reached its greatest efficiency in personal identification and application in forensic science. China forensic DNA database has already gathered tens of millions of data on parameters including, but not limited to, the DNA profiles, information related to the time, space, means, type of committed crime, and the residence, nationality, and individual behavior of the suspect. With the growing needs of public security, the database is still growing rapidly.

This paper discusses the combination of China forensic DNA database and technologies such as artificial intelligence and data mining based on big data. Through extensive analysis of the data from the DNA database, this paper seeks to uncover the underlying relation among data points in this database, and the relation between gene and social information. This paper also fully discusses the application potential of China forensic DNA database.

## 1. Analysis of big data attribute in China forensic DNA database

As of now, China forensic DNA database provides comprehensive coverage of information categories, contains enormous amount of data, and is still growing. The database mains high collection, updating, and processing speeds and efficiency. This database has great potential and contains rich data with great unrealized values. China forensic DNA database has played an important role in forensic science and have a great potential value. To a certain extent, it has the characteristics of big data such as Volume, Velocity, Variety, Value Veracity, Complexity and so on. First, the data of China forensic DNA database covers all geographic locations in China and all ethnicities. The collected information include nationality, ethnicity, family name and individual behavior of the suspect from all provinces and autonomously and ethnicities in China. The information has an unparalleled advantage. Secondly, China forensic DNA database also contains massive, multi category DNA profiles, such as Autosomal STR, Y-STR, mt-DNA, X-STR and SNP. The information has the characteristic of big genomic data, and gradually form a grid space which is the full coverage genes in China DNA database. Finally, China DNA database also stores various kinds of information related to the time, space, means, type of committed crime and action trail of the persons involved in the crime. The multi-dimensional information is the big data of social information.

## 2. Artificial intelligence analysis of forensic DNA database

In recent years, artificial intelligence (AI) and database technology are interpenetrating and merging with each other, which is bringing broader application prospects to the big data. The combination of forensic DNA data and AI will bring about a revolution in China forensic DNA database application.

Presently, the development of China forensic DNA database is mainly manifested in

horizontal development and is insufficient in depth, that is, the total amount of data is increasing, which is showing an explosive trend, and the data model still stays on the single structure of point to point search. When AI technology is introduced into China forensic DNA database, we can combine them together to realize the intellectualization of DNA database and to implement intelligent DNA database.

The intellectualization of China forensic DNA database is to treat the database as an AI system, which can be intellectualized by enabling retrieving, structure and association in the underlying mechanism of the database, so as to achieve efficient retrieval and matching in the big data.

Intelligent DNA database requires expanding of the existing database functions by enabling deductive, reasoning and intelligent analysis. By fully implementing the cloud computing technology in the DNA database, it is possible to develop a new mechanism that combine distributed computing, grid computing and parallel computing and can be used for knowledge acquisition, representation, updating and reasoning. It can play the role of multi-functional perception technology.

### 3. Research on comprehensive application of multiple genetic markers in China forensic DNA database

The revolutionary progress following gene sequencing has led the world to the era of big genetic data. With the development of DNA detection technology and big genetic data, a variety of genetic markers have been introduced into the field of forensic science, and the genetic markers in China forensic DNA database are gradually enriched. At present, China forensic DNA database contains a variety of DNA data profiles, such as Autosomal STR, Y-STR, mt-DNA, X-STR, SNP and so on. Based on the big data, China forensic DNA database will be able to integrating the functions of both genetic markers data storage and application. Through analysis of sex chromosomes genetic markers data, we can enhance identification effectiveness and assist in reducing investigation scope. For example, the paternal family by Y-STR data and the maternal family by mt-DNA data of the detected samples can be identified. The Facebook analysis can be realized by SNP data, and the locality of the detected samples can be predicted by gene frequency. By associating these data with multidimensional data structure, we can analyze complex relatives (such as single parent, brother, sister and so on) intelligently to some extent, so as to obtain important information related to crime, such as family, social relations and activity area.

The application of China forensic DNA database based on the big genetic data can form a fusion of data real-time collision modes, such as data matching, personal identification, identification of paternity, racial region and DNA characterization. According to the theory of big data, data matching and retrieval of DNA database with multiple genetic markers should focus on analyzing all the data related to a certain individual, instead of concentrating on only a small number of data samples. It should be focused on the integrity and regularity of the data, and no longer only pursue the accuracy of the data. It should be focused on the relationship between case, try to find information complete association, instead of deliberately seeking single cause and effect relationship.

### 4. Research on application of data mining technology in China forensic DNA database

The mining of big data is the inevitable outcome of information development to a certain extent, and it is the advanced stage of utilizing accumulated data. Data mining can find useful concepts, rules, rules, patterns and other useful knowledge from a big data through the search. The introduction of data mining into China

forensic DNA database can enable the analysis of complex types of data, such as DNA data, biological characteristics, personnel information and behavior, case characteristics and so on. At present, the greatest advantage and use of China forensic DNA database is individual identification. DNA data in the database is based on STR data, and the main analysis algorithm is matching and relative and relationship retrieval based on STR data. This application model has not fully explored the value of the DNA database. On the one hand, before the individual identification matching results based on DNA data are formed, the information about cases, evidence of the scene, and suspects are largely missing. On the other hand, the result of DNA database matching is usually one to one about two samples. The application is always isolated and not related.

Data mining technology can be used for in-depth analysis of data in DNA database.

1. Through the methods of association rules, clustering and SWOT analysis, the time, space, category, means related to crime, the regional and ethnic information of the suspect and the multi category of DNA gene data in the DNA database are analyzed and studied. The data model will be built which is used to show the dynamic relationship between high risk crime and time, space, population and other dimensions.
2. Through the methods of classification, correlation grouping, and clustering analysis, we can study the features of typical crimes in view of the crimes which are high incidence, obvious professional characteristics and great influence on social stability. The data model will be built which is used to show the distribution, evolution and prediction of typical criminal behavior in time and space.
3. Through the methods of classification, prediction, correlation grouping and clustering, we can study the activity track of the suspect in the DNA database. The data model will be built which is used to show the correlation of individual genetic data and behavior law.

The development capacity of a national forensic DNA database reflects the level of forensic science and the impartiality of justice in a country. The application of the forensic DNA database based on big data is a new technology with broad application prospects and challenges. The introduction of information technology, such as AI and data mining, is a direct result of big data and the open thinking of information. It is also an option for continuous progress, self-improvement and self-challenge of the forensic DNA database. China forensic DNA database based on big data which has a variety of genetic markers will show a kind of group relationship behavior. Through technical means of AI such as information value mining, human behavior analysis, semantic identification, intelligent analysis and so on, the relationship between the objective existence and the potential law can be found out from the various genetic data and the numerous data. It fully displays the potential and value of the application of the forensic DNA database in the field of criminal law research, crime dynamic analysis, public safety management decision and so on.