ISSN 2278-5612

A Collective Algorithmic Approach- For Enhanced DNA Database Security

Vijay Arputharaj J

Research Scholar, Department of Computer Science,
Karpagam University, Coimbatore, Tamil Nadu, India

phdvij@gmail.com

Dr.R.Manicka Chezian

Associate Professor, Department of Computer Science,
NGM College, Pollachi, India-642001

chezian_r@yahoo.co.in

ABSTRACT

The proposed method is a mixture of several security methods namely digital authentication tag along with the data mining in the DNA database. Data mining in the area of human genetics, an important goal is to understand the mapping relationship between the individual variation in human DNA sequences and variability in various algorithms for database security issues, for mutation susceptibility and parental identification differences. This paper primarily deals with the advancement of genetic algorithm with proper security features in DNA Databases and it enhances the special features in DNA database security. Several security methods include encryption algorithms, higher, not as much of multifaceted with trouble-free to apply in DNA Databases, used for protected database. The Reverse Encryption algorithm to protect data, Advance Cryptography algorithm to resist data, also Advanced Encryption Standard (AES) is most preferable for security in DNA databases.

Indexing terms/Keywords

DNA Database, Genetic Algorithm, Database Security, Security Algorithms.

Academic Discipline And Sub-Disciplines

Computer Science and Bioinformatics interdisciplinary

SUBJECT CLASSIFICATION

DataBase Security

TYPE (METHOD/APPROACH)

Literary Analysis and Experimental Analysis

Council for Innovative Research

Peer Review Research Publishing System

Journal: International Journal of Management & Information Technology

Vol.4, No.1
editor@cirworld.com
www.cirworld.com, member.cirworld.com

1. INTRODUCTION

The Human Genome Project is a worldwide scientific study mission with a main aim of formative the succession of chemical base pairs which structure DNA, also to identify and map the genes of the human genome from the corporeal and serviceable position. A DNA database or DNA databank is a database of contains all DNA data. A DNA Databank can be used in the analysis of parental comparison, genetic diseases, genetic fingerprinting for criminology, genetic genealogy etc. The security to the databases is often referred to as ultimate contour of protection. Database Security is a significant part of numerous applications where much sensitive information is to be stored in databases. When the data to be stored becomes too huge then the applications and its developers opt for using a database so as to make the application simple. These types of data are to be secured or must be kept away from hackers. Database security requires knowledge of both database management and security. There are various types of attacks that can occur to a database like database injection, referential integrity etc. The method of encrypting the database can prevent the sensitive data from wrong hands in almost all kinds of attacks. Cryptography is a method of securing data either over the network or in any stand alone device.

The goal of data mining in DNA Database is to check some possible combinations of DNA sequences and to generate a common sympathetic code or algorithm to formulate the sequence on mutations. Several visualizations and data mining techniques are already available, and they are used to validate and attempt to discover new methods for differentiating DNA sequences or exons, from non-coding DNA sequences or introns. Since the data mining is the best technique to analyze and extract the data, it is also helpful to formulate the common algorithm.

Data mining in the area of study on human genetics, an important goal is to understand the mapping relationship between the inter-individual variation in human DNA sequences and variability in disease, mutation susceptibility. In lay terms, it is used to find out how the changes in an individual's DNA sequence affect the risk of developing common diseases and mutations. This investigation also helps in parental identification algorithms for DNA sequences, genome expressions. Data mining, data extraction techniques are used to understand the need for analyses of large, complex, information-rich data sets in DNA Sequences.

Regulation of gene expression includes the processes that cells and viruses use to regulate the way that the information in genes is turned into gene products. An important challenge in use of large scale gene expression data for biological classification occurs when the expression dataset being analyzed involves multiple classes. To overcome this kind of problems data mining is used.

2. PROBLEM STATEMENT

2.1. Establishing a forensic DNA Profile/database

DNA profiling refers to the identification of particular parts of a person's DNA molecule. It is a technique which enables scientists to compare two biological samples and to determine the likelihood that these samples originated from the same individual. Because DNA is the same in all cells of the body, DNA profiles extracted from different samples at different times and in different places can be compared to determine whether they have come from the same person. If human biological samples are found at a crime scene, DNA profiling can determine whether a suspect could be a possible source of a sample. The supremacy of a forensic DNA database is, it can perform a superior role in the inquiry of crimes by connecting the DNA profiles from crime associated genetic sketch material to every supplementary and for the possible contributor (or their associations) of that genetic trace material.

More than the history of past 15 year's forensic DNA databases have established to be very powerful in this respect. In spite of this success not all ENFSI member countries have a DNA database yet.

The Council of the European Union invited Member States to consider establishing DNA Databases back in 1997. DNA database is a repository of DNA profiles generated from biological samples, which can be electronically stored for comparison with profiles generated from material found at the scene of a crime.

2.2. The Role of data mining in database security

Data mining is the practice of automatically analyze and extract meaningful and previously unknown pattern or knowledge form data stored in a large database. Data mining has also been defined as "The nontrivial extraction of implicit, previously unknown, and potentially useful information from data" and "The science of extracting useful information from large data sets or databases". Nowadays, there are many different models and algorithms used to perform data mining. Some of the examples are Decision Trees, k-nearest neighbor classification, and neural network. However, in the real life, many problems cannot be solved in polynomial amount of time using conventional methods. Under these situations, Genetic Algorithm is introduced to find the optimized solution.

2.3. The Role of Cryptography in database security

Classical database security relies on many different mechanisms and techniques, including access control, information flow control, operating system and network security, prevention of statistical inference, data and user authentication, encryption, time-stamping, digital signatures, and other cryptographic mechanisms and protocols. These methods together have the required mechanism to tackle any kind of attack on the database, since it uses all the techniques. But at the same time it may not be user friendly and it is time consuming because of the number of security measures that are to be satisfied. These type of techniques also require communication bandwidth so that the data transfer occurs between the database server and the client.

2.4. Data Mining using Genetic Algorithm

Genetic Algorithm derives from its similarity to the process of natural selection. Let the problem be to find a solution to a problem that would be the most optimal from the point of view of a certain criterion. And let each solution be exhaustively described by some set of numerical or non-numerical parameters. For example, if the task is to select a fixed number of market parameters influencing the market performance the most, then the names of these parameters comprise such a descriptive set. One can think of this set as of a set of chromosomes determining qualities of an "organism" - a solution of the problem. Following this analogy, values of parameters determining a solution correspond to genes. A search for the optimal solution is similar then to the process of evolution of a population of organisms, where each organism is represented by a set of its chromosomes

2.4.1 Uses of Genetic Algorithm

There are more reasons for preference using genetic algorithms in general in contrast to other techniques. One of them is its robustness and ability to work on large and “noisy” datasets, they perform global search of the solution space in comparison to most other algorithms that use greedy search, coping well with attribute interaction. Owing to all possible modifications and parallel approaches to genetic algorithms, the scalability of these algorithms can be achieved. Beside robustness, this characteristic is of great importance in data mining. Moreover, genetic algorithms have high degree of autonomy that enables discovery of knowledge previously unknown by the user. However, a drawback of genetic algorithms is the necessary frequent evaluation of individuals, i.e. possible solutions of the task, against the dataset. If we have *n* individuals and *m* attributes of the algorithm, the number of evaluations of individuals will be *nm*. If the datasets in data mining tend to have gigabytes and terabytes of data, evaluation of individual against the dataset is the most time consuming operation of the algorithm.

3. PROBLEM IMPLEMENTATION

The method used for insertion is explained below and works as following progressions

3.1. Digital implication

This is a basic security measure that is available in most of the systems. This login feature has a biometric technique with a user name and password. The feature has been designed in such a way that a particular special security implication and which consists of available border level of security features.

3.2. Data improvisation using DM technology

The data for the database is accepted. The type of encoding has to be specified For insertion the values of the fields in database are given. For updating, initially the primary key value and the encoding type are given. If it is found to be valid then the other field values are obtained. The encoding type of a row cannot be changed while updating. In updating if the encoding type or the primary key value of the database is specified wrongly then the system automatically comes out of application.
Formal knowledge discovery in database process (KDD) will be adopted to perform:
1. Learning the application domain to extract relevant prior knowledge and goals of application
2. Creating a target data set: data selection
3. Data cleaning and preprocessing:
4. Data reduction and transformation:
5. Choosing data mining approach
   - Classification
   - Regression
   - Association
   - Clustering
6. Choosing the mining algorithm(s)
7. Data mining: search for patterns of interest
8. Pattern evaluation and knowledge presentation
   – Visualization, transformation, removing redundant patterns, etc.
9. Use of discovered knowledge

3.3. Data improvisation towards Database security

All the data given in the above step are given to the encryption algorithm for encryption. These are the data that are to be encrypted before adding to the database.
The encryption process has the following processes.
Cryptography is usually referred to as "the study of secret", while nowadays is most attached to the definition of encryption. Encryption is the process of converting plain text "unhidden" to a cryptic text "hidden" to secure it against data thieves. This process has another part where cryptic text needs to be decrypted on the other end to be understood.
In the broad meadow of cryptography, encryption is the procedure of indoctrination letters (or information) within such a method that hackers cannot understand writing it, other than that approved parties only can used it.
In an encryption scheme, the memorandum or information, it is also called as plain text; this text is encrypted using an encryption algorithm, turning it into an unreadable cipher text. This is usually done with the use of an encryption key, which specifies how the message is to be encoded. After that decryption is also done by the authorized party.
Encryption is a method of hiding data so that it cannot be read by anyone who does not know the key. The key is used to lock and unlock data. To encrypt a data one would perform some mathematical functions on the data and the result of

these functions would produce some output that makes the data look like garbage to anyone who doesn't know how to reverse the operations.
The Advanced Encryption Standard (AES) is a measurement for the encryption of electronic records which is conventional scheme by the U.S.National Institute of Standards and Technology (NIST) in 2001,

BASE STEPS IN THE SECURED DB

1. KeyExpansion—round keys are derived from the cipher key using Rijndael's key schedule.
2. InitialRound
    1. AddRoundKey—each byte of the state is combined with the round key using bitwise xor.
3. Rounds
    1. SubBytes—a non-linear substitution step where each byte is replaced with another according to a lookup table.
    2. ShiftRows—a transposition step where each row of the state is shifted cyclically a certain number of steps.
    3. MixColumns—a mixing operation which operates on the columns of the state, combining the four bytes in each column.
    4. AddRoundKey
4. Final Round (no MixColumns)
    1. SubBytes
    2. ShiftRows
    3. AddRoundKey

THE GENETIC OPERATORS FOR ENHANCED SECURITY BASED DATA MINED ALGORITHM: These operators permit GAs en route in favor of explore the search space. But, operators characteristically comprise disparaging as well as beneficial possessions. They must be tailored to the predicament.
Crossover: We use in database a Subset- Size Oriented Common Feature Crossover Operator (SSOCF) that maintains practical enlightening building blocks as well as constructs sour element's that contain the similar allocation than the base. Sour elements are kept, merely if they on top form improved than the slightest good quality entity of the dataset element residents. Special Features collected by the base elements are kept by sour elements and the non shared features are inherited by sour elements corresponding to that particular numberth base with the probability features.
Transformation: The transformation is an operative element which agrees to assortment. Throughout the transformation phase, that chromosome has a likelihood $n_{dna}t$ to transform. If a chromosome is selected to transmute, that particular element is chosen arbitrarily a number n of elements to be flick then $n_{dna}t$ elements are flicked automatically for algorithmic calculation. To generate a huge assortment, we set $n_{dna}t$ around 10% and n € [1,5].
Selection: The implementation with database in a twofold contest mixture. Selection event which holds m contests to choose j individual elements. Each contest has of its own variety of set essentials of the inhabitants and choosing the best one with a likelihood p €¸[1,5].

5. ADVANTAGES AND DISADVANTAGES

The advantage of using genetic algorithm is that it doesn't have to know any rules of the problem in advance – the rule will can be found through evolution. This is very useful for very complex and loosely defined problem. Also, with a well defined fitness function and carefully chosen attributes, genetic algorithm can perform much faster than other algorithm such as the linear method.
The drawback of genetic algorithm is that the definition of the fitness function can be very complicated sometime. The fitness function may affect the performance of the process significantly if the complexity of the fitness function increase. It is because the fitness function is used to compare every element in the sample population to every data in the training data set. Sometimes an acceptable solution cannot be derived even after countless iteration if the genetic operators are wrong chosen**.**

6. CONCLUSION

In this paper, the basic knowledge of combination of several techniques like data mining, genetic algorithm, database security algorithms were covered. In addition, a model using genetic algorithm for data mining was developed. The model was then applied on a real world secured database in order to solve a specific problem. The preliminary simulation of the process showed that the derived model can progressively increase the coverage of the rule. In the future work, the algorithm derived in this paper will be implemented into program using enhanced bulk databases. Beside, the study will be focus on applying genetic algorithm on invalid faction. Finally, it will compare with conventional data mining technique in order to find the benefit by using genetic algorithm with a special combination design of security algorithms.

ACKNOWLEDGMENTS

Our thanks to the experts who have contributed towards development of the template.

REFERENCES

[1] W. Frawley and G. Piatetsky-Shapiro and C. Matheus, "Knowledge Discovery in Databases: An Overview". AI Magazine, Fall 1992, Pages 213-228.
[2] D. Hand, H. Mannila, P. Smyth, "Principles of Data Mining". MIT Press, Cambridge, MA, 2001. ISBN 0-262-08290-X
[3] http://www.megaputer.com

[4] Richard J. Roiger and Michael W. Geatz, “Data Mining: A Tutorial-based Primer” Pearson Education, Inc, 2003. Page 90. ISBN 0-201-74128-8
[5] Li Lin, Longbing Cao, Jiaqi Wang, Chengqi Zhang, "The Applications of Genetic Algorithms in Stock Market Data Mining Optimization", Proceedings of Fifth International Conference on Data Mining, Text Mining and their Business Applications, 2004
[6] P. Seligman, S.W. Smith, "Detecting Unauthorized Use in Online Journal Archives: A Case Study." Proceedings of the IADIS International Conference WWW/Internet 2004. Volume 1. 209--217. October 2004.
[7] Michelle Finley, “Smart Methods to Spot Fraud”, WIRED News (April 3, 2000).
[8] P. van der Putten and M. van Someren (eds). CoIL Challenge 2000: The Insurance Company Case. Published by Sentient Machine Research, Amsterdam. Also a Leiden Institute of Advanced Computer Science Technical, Report 2000-09. June 22, 2000.
[9] http://www.genetic-programming.com/gpquadraticexample.html
[10] John Koza, “Genetic Programming: On the Programming of Computers by Means of Natural Selection”; Page 117. ISBN 0262111705
[11] Vijay Arputharaj J and Dr.R.Manicka Chezian (2013), Data mining with human genetics to enhance gene based algorithm and DNA data base security, International Journal of Computer Engineering and Technology, PP 176-181,
[12] G.W. Moore. (2001). Cryptography Mini-Tutorial. Lecture notes University of Maryland School of Medicine. Internet: http://www.medparse.com/whatcryp.htm [March 16, 2009].
[13] T. Jakobsen and L.R. Knudsen. (2001). Attack on Block of Ciphers of Low Algebraic Degree. *Journal of Cryptography*, New York, *14*(3), pp.197-210.
[14] N. Su, R.N. Zobel, and F.O. Iwu. “Simulation in Cryptographic Protocol Design and Analysis.” Proceedings 15th European Simulation Symposium, University of Manchester, UK., 2003.
[15] Dr.R.Manicka Chezian, and Dr.T.Devi. “Termination of triggers in active databases” International Journal of Information Systems and Change Management, USA, Vol-5, No-3 PP 251-266, 2011
[16] Dr.R.Manicka Chezian, and Dr.T.Devi. “A new algorithm to detect the non termination of triggers in active databases” International Journal of Advanced Networking and Applications, Vol-3, Issue-2 PP 1098-1104, 2011
[17] Dr.R.Manicka Chezian, and P.M.Nishad “A vital approach to compare the size of DNA sequence using LZW with fixed length binary code and tree structures” International Journal of Computer Applications, Vol-3, No-1, PP 7-9, 2012
[18] Dr.R.Manicka Chezian, and C.Bagyalakshmi “A survey on cloud data security using encryption technique” International Journal of Advanced Research in Computer Engineering and Technology, Vol-1, Issue-5,PP 263-265, 2012

Author’ biography with Photo

Author Mr.Vijay Arputharaj J is a part time PhD research scholar of Karpagam University, Coimbatore, India. Curently he is working as a Assistant Professor at VLB Janakiammal College of Arts and Science, Coimbatore, he has attended and published various conferences and research journals.

Co-Author Dr.R.Manicka Chezian is Associate professor at Dr.NGM College, Autonomous, Pollachi, India. He is guiding a number of PhD and MPhil Scholars under his leadership skills. He has attened more than 10 international and national coferences and he has published more than 20 journals in international and national journals.