ISSN No. 0976-5697

Volume 5, No. 6, July-August 2014

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

A Dynamic DNA for Key Based Cryptography Approach

Yesha Pruthi [*1], Sunita Dixit[2]

M.Tech. Student, CSE Dept.[1], Assistant Professor in CSE Dept. [2]

PDM College of Engineering for Wome,India

Abstract— Cryptography is always taken as the secure way while transforming the confidential information over the network. But over the time, the traditional cryptography approaches are been replaced with more effective cryptographic systems such as quantum cryptography, biometric cryptography, geographical cryptography and DNA cryptography. The presented work is about to defined a DNA cryptography scheme in which DNA concept will be implemented at two stages. At the initial stage, the DNA sequence will be used as the key to perform the cryptography and later on the cryptography itself will be done using the DNA coded approach. The presented work will be implemented on text cryptography. At the earlier stage, the system will accept the DNA sequence as the input to generate the key. The DNA sequence will be identify the maximum frequency DNA pattern. This sequence pattern will be used as the key to the cryptography system. At the later stage, the DNA encoded dictionary will be defined for cryptography. Now the input text will be coded using this DNA coded dictionary. The presented work will also optimize the process of DNA sequence search mechanism so that the effectiveness of the cryptographic algorithm will be improved.

Keywords: DNA cryptography, dynamic key, coded dictionary, tandems

I. INTRODUCTION

Data Encryption enables the information security while performing the private communication in public network. Cryptography is the encoding approach used to convert the raw information in encoded form so that the data integrity over the network is improved. When the data is communication in an open network, the data is having number of security issues in terms of information attack performed by some attacker or intruder. Because of this, there is requirement of some more reliable way of information transmission over the network. These ways of information security includes cryptography, steganography, message digest etc etc[2]. These all approaches modify or secure the information in some way. One of the most traditional way to secure the information is cryptography. Cryptography actually modify the information itself in some encoded format the can be decoded back to original form. It is some kind of handshaking mechanism in which, the sender sends the information in some encoded format and the information can be retrieved only by the receiver back. If some intruder gets the information packets even then information is not readable. In such case, the security of information depends on the encoding mechanism. More secure and reliable the encoding mechanism, more secure information transfer is possible over the network[1].

In case of electronic data communication the requirement of such cryptographic approach is more critical. Different kind of secure communication or digital transaction increased the need of cryptography approaches. Such as credit card payments, e transactions requires more concern of user to use the secure communication. Today emails and SMS are also communicated in secure way. Many of the service provides also available web information security in secure means and present the information under the trust level. A trustful communication medium is more reliable to provide the effective communication over the network. This kind of secure communication is not only required for public networks but while performing the offline data transmission, the cryptography approaches are more beneficial. These approaches also secures the information from social hacking or the attack done by some known person.

Here to secure the information, the cryptography mechanism basically requires two main components called cryptographic algorithm and the secure key[5]. Cryptographic algorithm is the approach applied to encode the information and the key is the actual password information used by the sender to encode the information. A cryptography algorithm accepts the raw textual information and the sender key as input and perform the encoding mechanism. After this encoding mechanism, the cipher text is obtained. This encoded information travels over the network. As the receiver or some intruder gets this encoded information, they requires the knowledge about the decoding algorithm and the relative decoding key. If the algorithm or key is not available it is not possible to retrieve the data back from encoded form.

Cryptography enables the data communication between two persons or between a group to provide secure communication over the network[2]. This kind of communication not only prevents the unauthorized communication over the network but also maintains the data integrity. The authentication is provided by digital signature or digital certificates[1].

A. Types of Cryptography:

At the basic level cryptography approaches are divided in two main categorized based on number of keys involved in communication. These approaches are called private key cryptography and public key cryptography.

SECURING DATA THROUGH A CRYPTOGRAPIC PROCESS

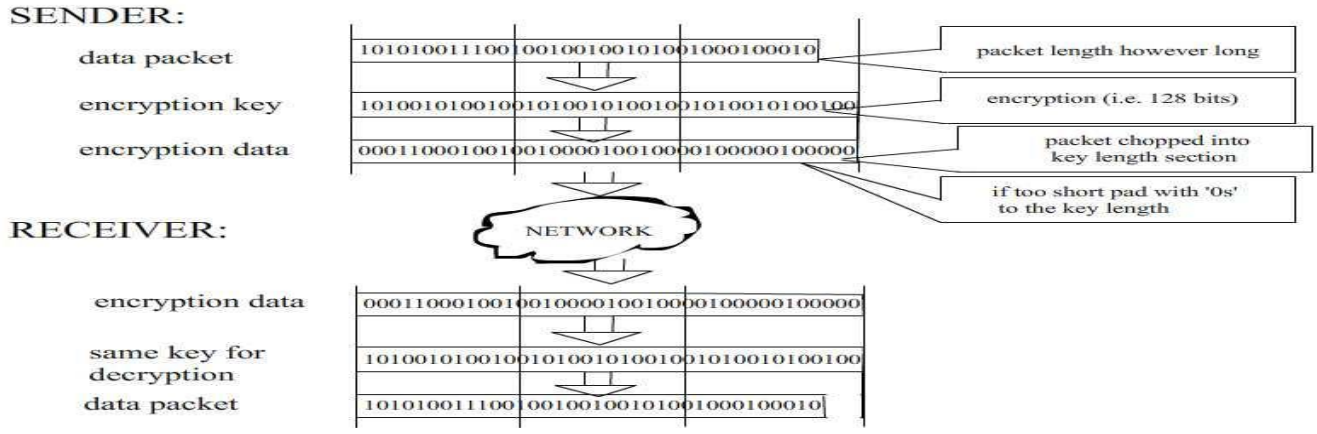

Figure: 1

B. Symmetric Key Cryptography:

This kind of cryptography approach is also called private key cryptography. As the name suggested in this cryptography approach only single key is involved to enable the encoding and decoding process. It means same key is used to perform data encryption and to retrieve the data back from cipher text. At the earlier stage, this kind of cryptography not looks more stronger as the complete security depends on single key. But there are number of cryptography algorithms comes under symmetric key cryptography that increases the data integrity by using the larger key size and number of encoding level in the algorithmic approach[4]. This is the most traditional type of cryptography, in which the key information is common for both sender and receiver. In such system, the sharing mechanism of key requires some effective approach. Single key is here defined to perform data encoding and decoding. This single key is able to provide the reliable communication over the network. The cryptographic mechanism supported by this approach is shown in figure 1.2

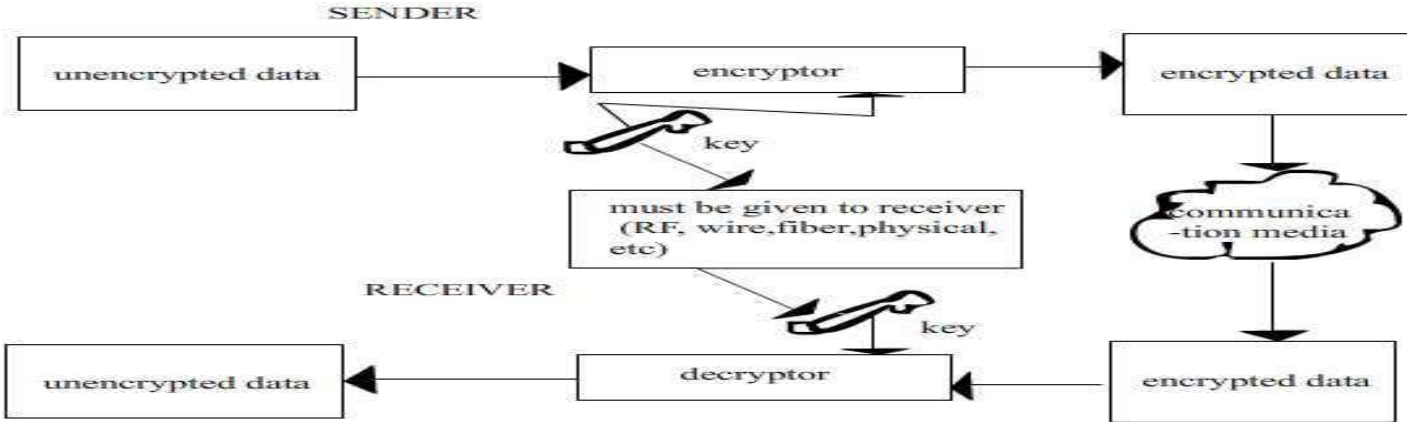

Figure 1.2 Symmetric Key Cryptography

As figure shows, the sender is having the raw information to transfer over the network. This raw information can be available in different media types[4]. These media types can be textual information, image, videos etc. To perform the cryptography, the approach requires some algorithm called encrypt or and the key. This key is symmetric key shared between the sender as the receiver. As the cryptography algorithm is applied, the information is encoded to the cipher data form or called encoded information. Now this encoded information is transferred over the network. As the receiver receives the information it is in encoded form. Now the decrypt or is applied here to get the actual information back. The decryptor uses the decoding algorithm and the same symmetric key to get the information data back[8]. The security of these kind of algorithm depends on three main vectors called cryptography algorithm, key size and way to share the key. There are number of symmetric key cryptography algorithms such AES, DES (Digital Encryption Standard), Triple DES, AES (Advanced Encryption Standard) etc. These algorithm provides the high level information security[8].

C. Public-key cryptography:

This cryptography approach is also called asymmetric cryptography approach. According to his approach, encryption and decryption is performed by two different keys. According this cryptography approach. As the encryption process begins, instead of generating single key, two keys are generated called Public key and private Key[4]. The generator A keeps the private key with himself and distribute the public key to all users that can send information to it. Now, as some user B want to send information to the user A. In such case, user B will use the

public key of User A to perform the encoding process. Here the cryptography will be performed using public key of receiver. Now after the encoded process, the cipher information is transferred to the receiver A. As receiver receive this information, the decoding process is performed using private key of User A[4]. This decoding process is able to get the information back in its original form. Complete cryptography process is here shown in figure 1.3.

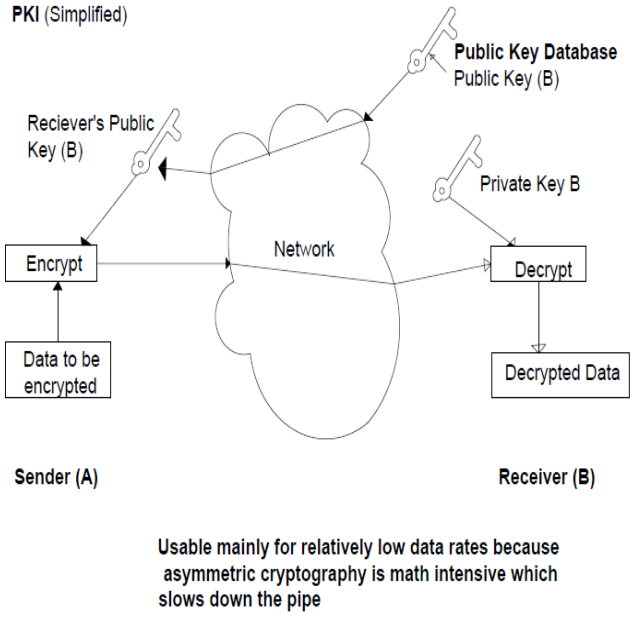

Figure 1.3: Public Key Cryptography

As the process shows, the public key cryptography is more secure to perform data communication over the network. This cryptography approach uses different keys to enable the communication over the network. The algorithms available for public key cryptography includes RSA. The security of this algorithm depends of the algorithm type and the key size.

II. DNA

Today, the scope of computer science is increasing in different application areas because of its auto gaining and evolutionary features. One of such application area in which, the use of computer science has provided the outstanding and evolutionary results. The application area is bioinformatics that combines the biological information with information technology[7]. The biological information is here plasticized to generate new genomic revolution. This information is available in the form of genetic sequence.

The main objective of bioinformatics is to identify the relationship between large dataset group and sequences. This dataset groups are available in the form of protein, acider sequences, structures etc. This structural information can be presented in different forms. Some of these forms includes DNA, RNA, Protein, Genetic Codes etc. All the living things and organisms are defined by a functional cell unit that represent it as the alive. In each cell, number of processes resides that are controlled by protein. These proteins are composed using the molecules of amino acid. The component of amino acids, enzymes are useful for the construction of DNA. These component amino acids, enzymes which break down fat molecules and enzymes that allow ingested nucleic acid to be reused for the construction of DNA[9].

DNA itself defines the instruction code for genetic and by using it the protein structure of any living thing can be constructed or recomposed. Each DNA sequence is based on four different bases called Adenine (A), cytosine (C), guanine (G) and Thymine (T). DNA sequence study is helpful to do the structural change or property change or behaviour change for a particular living thing[3]. To identify the characteristic match or to identify the similarity between two living things in terms of characteristics or the functionality or behaviour, it is required to analyse them respective to their DNA sequence[15]. Each DNA sequence is defined as the large information group that contains all kind of information about the living things. Such as if we take the example of wheat, it contains the information about the wheat colour, smell, quality etc. If we have to find a particular quality of wheat among the wheat samples, the DNA sequence match for the particular pattern can be performed. Each pattern in DNA sequence represents the existing or non-existence of some characteristics or the behaviour. But the identification of these patterns over the DNA sequence is a challenging task. In this present work, the main focus is to perform the identification of some of such patterns over the DNA Dataset Group[10].

The presented work is focused on same concept of bioinformatics sequence to perform a pattern search over the DNA sequence. The work is divided in two main stages, The first stage is about to generate the pattern sequence itself by performing the study on DNA sequence.

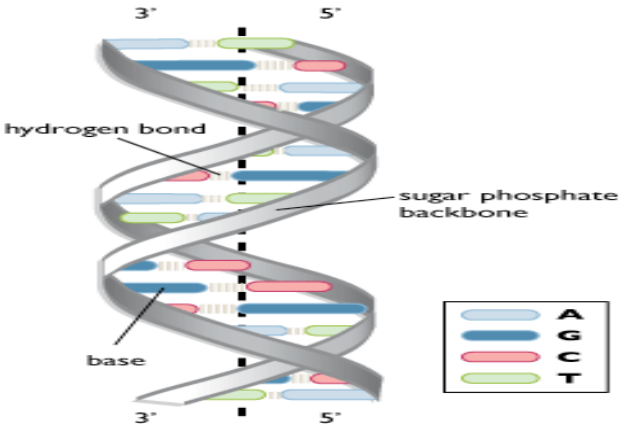

Figure 2. Double helical DNA structure showing its basic components

III. DNA CRYPTOGRAPHY

Cryptography is the most significant component of the infrastructure of com- medication security and computer security. However, there are several latent defects in a number of the classical cryptography technology of recent cryptography-such as RSA and DES algorithms - which are broken by some attack programs. DNA cryptography and knowledge science was born after analysis within the field of DNA computing field by Adleman; it's It is an field and has become the forefront of international research on cryptography[11]. Several researchers from global countries have done an outsized range of studies on DNA cryptography. Adleman suggested a proof-of-concept use of DNA as a type of computation which solved the seven-point Hamiltonian path drawback.

Since the initial Adleman experiments, advances are created and varied Alan Turing machines are verified to be constructible.DNA Cryptography relies on biological problems: in theory, a deoxyribonucleic acid system won't only has constant computing power as a contemporary system however It has have a efficiency and function traditional ancient computers cannot match[13]. First, deoxyribonucleic acid chains have a really large scale of parallelism, and its computing speed may reach 1 billion times per second; second, the deoxyribonucleic acid molecule - as a carrier of information - encompasses a huge capacity. It appears that one trillion bits of binary information may be stored in one cubic decimetre of a deoxyribonucleic acid solution; third, a DNA molecular system has low power consumption, solely up to billionth of a traditional system[11]. The research of DNA cryptography is still at its initial stage, Viviana Risca's a computer scientist proposing "Hiding messages in DNA mi- crodots". DNA Computing relies on biological problems: in theory, a deoxyribonucleic acid system won't only has constant computing power as a contemporary system however It has have a efficiency and function traditional ancient computers cannot match. First, deoxyribonucleic acid chains have a really large scale of parallelism, and its computing speed may reach 1 billion times per second; secondly, the deoxyribonucleic acid molecule - as a carrier of information - encompasses a huge capacity[14]. It appears that one trillion bits of binary information may be stored in one cubic decimetre of a deoxyribonucleic acid solution; third, a DNA molecular system has low power consumption, solely up to billionth of a traditional system[9].

IV. PROPOSED WORK

The presented work is about to defined a DNA approach to perform the encryption on text. The work is divided in two layers to achieve the efficiency as well as the reliability. In first layer, the generation of the dynamic key based on DNA sequence will be done. In the second layer, the actual DNA dictionary based encryption will be done so that encoded text will be obtained. The work will improve the reliability and the security of encoding process[17].

A. Objectives:

The objectives of the proposed work are listed here

a. The foremost objective of the work is to define a two stage DNA cryptography approach to encode the text.
b. The objective of the work is to define optimize approach to identify the dynamic DNA sequence that will be used as the cryptography key.
c. The objective of the work is to define a dictionary based mechanism to perform text encoding.
d. The objective of the work is to improve the reliability and security of communication system.`
e. The objective of the work is to implement the work in java environment.

B. Hypothesis:

The hypothesis is about the research questions that we have to achieve during the research work. A research work is about identifying the answers

a. Is the presented work is working for all kind of Text ?
b. Is it providing the robustness in terms of device ?

In this presented work a two stage, DNA cryptography approach is presented to text encryption. In this work, the DNA concept will be used for the key generation as well to encode the text. At the earlier stage, the dynamic DNA pattern will be identified over the sequence to generate the key to perform the encoding. Later on, the DNA code dictionary will be defined to perform the cryptography. The model of the presented work is given here under[13].

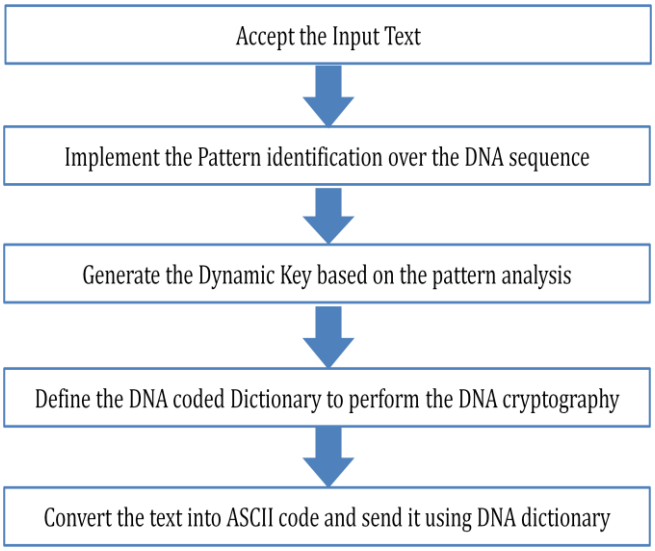

Figure: 3

C. DNA Sequence Mining:

The frequent pattern mining of the DNA sequence is an important mean to study the structure and function of the DNA sequence. In this paper, base on the characteristics of the DNA sequence, to propose the algorithm of JMPS(joined maximal pattern segment), which use of the maximal frequent pattern segments base on adjacent to the maximal frequent pattern mining, to improve the availability and efficiency of the DNA sequence data mining. DNA sequences use an alphabet {A, C, G, T} representing the four nitrogenous bases Adenine, Cytosine, Guanine and Thymine.

The Homo Sapiens (human) DNA sequence AX829174 starts with TTCCTCCGCGA and contains 10,011 characters[6]. The subsequence mining problem is of particular importance in computational biology, where the major challenge is to detect short sequences, usually of length 6- 15, that frequently occur in a given set of DNA or protein sequences.

These short sequences can provide clues regarding the locations of so called "regulatory regions," which are important repeated patterns along the biological sequence. The repeated occurrences of these short sequences are not always identical, and some copies of these sequences might differ from others in a few positions. The similarity metric which is used here could be complex—for example, when comparing protein, a similarity matrix like PAM or BLOSUM , may be used for comparing the "distance" between each symbol (protein) pair.

These patterns occurring frequently are called motif in computational biology[15].

We use this term to describe frequently occurring approximate sequences. Different similarity models require

different applications to suit the kind of noise that they deal with.

It is desirable for a motif mining algorithm to be able to deal with a variety of notions of similarity.

D. Research Design:

The proposed work is about to find the ten dam repeat patterns over the DNA sequence so that we can find the search the DNA pattern. In this work a two dimensional model or the structure is been represented in which the input DNA Sequence will be filled such way so that the search of any DNA pattern can be done accurately and efficiently. The presented work is defined as the algorithm that can be implemented and integrated with any text based application.

The Research Design is given as

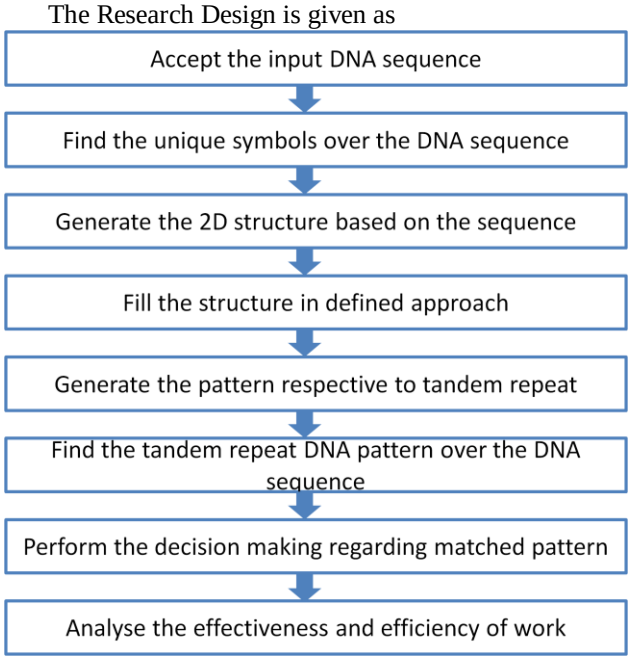

Figure: 4

E. Implementation:

Complete work of DNA tandem repeat sequence pattern identification work is divided in number of sub stages. These stages are presented in the form of separate algorithm. These algorithms includes the

(i) Generation of Frequency Matrix of DNA sequence alphabet

(ii) Search of a DNA pattern over the sequence (Single Alphabet, 2 Alphabet, Multiple Alphabet Sequence)

(iii) Generation of Tandem Repeat Sequence

(iv) Sequence mining of Tandem Repeat Pattern over the DNA Sequence.

V. CONCLUSION

DNA sequence mining is one of the most increasing bioinformation application and utility, that can be applied to perform the knowledge discovery over the DNA sequence[19]. These patterns represents the characteristics of a living thing so that the characteristic discovery and matching are the common operations of DNA sequence mining. In this present work, these two DNA sequence mining approaches are combined[16]. The work is divided in two main stages. In first stage, the DNA key generation was preformed based on DNA sequence analysis. The work is defined as the automated tool that will analyse the DNA sequence and identify all the possible patterns over the sequence. Once the key is generated, it is included in dataset in transformation form to encrypt the data. Later on the cryptography is applied over it perform the DNA cryptography[17].

VI. ACKNOWLEDGEMENT

I would like to thank for their paper ‘Naveen Jarold, P Karthigaikumar, N M Sivamangai, Sandhya , Sruthi B Asok, ”Hardware Implementation of DNA based Cryptography”,’ published in IEEE International Conference on Communication Technology.

VII. REFERENCES

[1]. Naveen Jarold, P Karthigaikumar, N M Sivamangai, Sandhya , Sruthi B Asok, ”Hardware Implementation of DNA based Cryptography”, Conference on Information and Communication Technologies@ 2013 IEEE

[2]. Haym Hirsh," Using Background Knowledge to Improve Inductive Learning of DNA Sequences", 1043-0989/94@ 1994 IEEE

[3]. Patrick Hoffman," DNA Visual And Analytic Data Mining", Proceedings of the 8th IEEE Visualization '97 Conference 1070-2385/97© 1997 IEEE

[4]. Xiong Wang," Finding Patterns in Three-Dimensional Graphs: Algorithms and Applications to Scientific Data Mining", IEEE Transactions On Knowledge And Data Engineering 1041-4347/02@2002 IEEE

[5]. Arvind Rao," A Clustering Algorithm for Gene Expression Data using Wavelet Packet Decomposition", 0-7803-7576-9/02@2002 IEEE

[6]. Jian Pei," MaPle: A Fast Algorithm for Maximal Pattern-based Clustering", Proceedings of the Third IEEE International Conference on Data Mining (ICDM’03) 0-7695-1978-4/03 © 2003 IEEE

[7]. Haixun Wang," A Fast Algorithm for Subspace Clustering by Pattern Similarity", Proceedings of the 16th International Conference on Scientific and Statistical Database Management (SSDBM’04) 1099-3371/04 © 2004 IEEE

[8]. Zonghong Zhang," Mining Deterministic Biclusters in Gene Expression Data", Proceedings of the Fourth IEEE Symposium on Bioinformatics and Bioengineering (BIBE’04) 0-7695-2173-8/04 © 2004 IEEE

[9]. Daxin Jiang," Cluster Analysis for Gene Expression Data: A Survey", IEEE Transactions On Knowledge And Data Engineering 1041-4347/04@2004 IEEE

[10]. Jagdish Chandra Patra," Neural Networks for Gene Expression Analysis and Gene Selection from DNA Microarray", Proceedings of International Joint Conference on Neural Networks, 0-7803-9048-2/05©2005 IEEE

[11]. Jin Pan," Efficient Algorithms for Mining Maximal Frequent Concatenate Sequences in Biological Datasets", Proceedings of the 2005 The Fifth International Conference on Computer and Information Technology (CIT’05) 0-7695-2432-X/05 © 2005 IEEE

[12]. Yu He," Mining Frequent Patterns with Wildcards from Biological Sequences", 1-4244-1500-4/07©2007 IEEE

[13]. Tae Ho Kang," Mining Frequent Contiguous Sequence Patterns in Biological Sequences", 1-4244-1509-8/07@2007 IEEE

[14]. Feida Zhu," Efficient Discovery of Frequent Approximate Sequential Patterns", Seventh IEEE International Conference on Data Mining 1550-4786/07 © 2007 IEEE

[15]. Xiaonan Ji," An efficient technique for mining approximately frequent substring patterns", Seventh IEEE International Conference on Data Mining - Workshops 0-7695-3019-2/07 © 2007 IEEE

[16]. Fabio Fassetti," Mining Loosely Structured Motifs from Biological Data", IEEE Transactions On Knowledge And Data Engineering 1041-4347/08@ 2008 IEEE

[17]. Quanwei Zhang," Genetic K-modes based DNA Splice Site Adjacent sequence_ Feature Analysis", Proceedings of the 7th World Congress on Intelligent Control and Automation 978-1-4244-2114-5/08© 2008 IEEE

[18]. Robertas Damasevicius," Analysis of Binary Feature Mapping Rules for Promoter Recognition in Imbalanced DNA Sequence Datasets using Support Vector Machine", 2008 4th International IEEE Conference "Intelligent Systems" 978-1-4244-1739-1/08© 2008 IEEE

[19]. Fang-Xiang Wu," On Determination of Minimum Sample Size for Discovery of Temporal Gene Expression Patterns", Proceedings of the First International Multi-Symposiums on Computer and Computational Sciences (IMSCCS'06) 0-7695-2581-4/06 © 2006 IEEE