# Data Encryption in Communication Using DNA Sequences

B.Bazli, D.Llewellyn-Jones, M.Merabti
School of Computing and Mathematical Sciences
Liverpool John Moores University, Liverpool, UK
{B. Bazli, D. Llewellyn-Jones, M. Merabti@ljmu.ac.uk}

**Abstract – cryptography is a field which makes the transmitted message unreadable to unauthorised users. In this work we take inspiration from DNA encryption schemes and use of biological alphabets to manipulate information by employing the DNA sequence reaction to autonomously make a copy of its threads as an extended encryption key. Information is converted from plain text to several formats and then follows the stages of protein formation from DNA sequences to generate an extended key using chemical property and attributes to be used in encryption mechanism. This technique will enhance the security of the encryption mechanism by substitution, manipulation, and complexity. Furthermore this technique can be used in many applications of information and communication systems as well as adding more complexity to existing encryption algorithms.**

***Keywords— DNA Sequence; Encryption; security; Bioinformatics; Communication***

## I. INTRODUCTION

Protection of data communication and sensitive information from unauthorised access has always been a challenging task. Several protocols and standards exist to safeguard communication networks and its assets, but intruders always find new tools and techniques to access information flows violating the privacy and security of those legitimate users. While protecting the information flow is a never ending task with ubiquitous use of the technology within personal, commercial, and technological environments, much of the attention has been paid to reactive protection of networks with known tools and techniques learned from previous incidents.

Process of converting messages from plain text to cipher text is called cryptography. Cryptography is a technique of achieving security for communications by encoding plain text messages to make it unreadable[1]. Encryption is a useful tool in protecting confidentiality and integrity of information. It is simply a technique for hiding the true meaning of the information from unauthorised users. The worst case of an attack within communication is complete control of the encryption system by illegitimate users. This happens by accessing the encryption algorithm to decrypt the data and access sensitive information. Cryptography relies on uncertainty in encoding the message to its cipher format. Redundancy in the known human languages [2] and limitations and flaws of the cryptography methods make them vulnerable especially to frequency analysis based attacks. A cryptanalyst can apply a frequency analysis based attack with the most repetitive letters, 'E' and 'I', to extract the message. With the entire precaution, security policy, and the complex algorithm, one thing is certain regarding the cryptosystems; if the attacker accesses the key that is used to encrypt the message, the message becomes readable.

DNA computing started by Adleman [3] trying to solve a small instance of the Hamiltonian path problem using parallel computing. DNA is considered as a medium for ultra-compact information storage, exceeding capability of conventional electronic media. A few grams of DNA may hold all data stored in the digital mediums in the world.[4]

DNA based bimolecular cryptography design is a technique which uses the huge parallel processing capabilities of bio molecular computation which convert short messages from hexadecimal and ASCII forms and then back to encrypt/decrypt the information. This has been used on different applications but we consider using this technique to safeguard sensitive information with the addition of the key generation technique from the bio-molecular properties of the DNA sequences.

The unique property of DNA encoding is used for computations, improve the security and encryption and to mitigate the flaws of the current security mechanism. In this work we take inspiration from DNA to manipulate information by employing the DNA sequence reactions to autonomously make a copy of its threads with a high fidelity for comparison. Furthermore we use chemical properties of the sequences as an indexing keys to authenticate the communications and hide information from exploitation techniques used by intruders specially frequency analysis based attacks. The DNA sequences and different stages from aligning the sequences with other biological languages and process of protein construction from the source property can be used to manipulate and encode data with inspiration from such methods. We take this into account to go through current security methods and encryption algorithms to propose a novel algorithm to enhance the security and complexity of an encryption mechanism.

The rest of the paper is organised as follows; section 2 describes DNA encoding mechanism in bio molecular computation using transcription and translation; Section 3 highlight related works and brief analysis of them; Section 4 describes the proposed algorithm through cooperative discussion; Section 5 discusses the proposed scheme using different resources; and section 6 concludes with summary of the paper.

ISBN: 978-1-902560-27-4 

## II. BIOMOLECULAR COMPUTATION

DNA is a sequence of nucleotides. The exact sequences of the nucleotides determine the code of each gene. DNA sequences represent biological information such as skin colour, weight, nose shape, eye, and hair as well as other features[5]. A DNA sequence is of a long molecule with four bases called nucleotides Adenine (A), Guanine (G), and Cytosine (C) and Thymine (T). DNA Sequences is succession of those letters that indicate order of nucleotides. Because of weak forces between the sequences, they pair as A-T and G-C. They form a chain around each other in the opposite direction to form a double helix. Although there are only four bases in the sequence, their arrangement in the long double helix is random and can be billions of combination of codons. This is how everybody in the world has different DNA. Such a capacity has created a field to mathematicians and cryptanalysts to explore the capability and functionality of the DNA sequences and bio molecular computation. DNA computing and its capability is used for parallel computing. The potential of DNA allows the researchers to solve numerous computational problems by parallel processing.

### *A. DNA Encryption*

DNA stands for Deoxyribonucleic Acid, a genetic material in human organs. The information in DNA is stored as code of four chemical substances namely; Adenine A, Cytosine C, Guanine G, and Thymine T. The order and sequences of these bases provide information about individuals such as peoples' name formed with alphabetical appearance[6]. This provides capacity and potential for many mathematical and statistical solutions dealing with data and provides naming, addressing and other functionality. The computational capability of DNA has been found by Leonard M Adleman[3]. DNA based bio molecular cryptography design is a library of one-time-pads assembled secretly in the arrangement of DNA strands which is used to encrypt or decrypt short messages[4].

The computation carried out using a DNA sequence is called DNA Computing. Diverse problems with significant storage capacity have been solved using parallel computing methods [8]. J.D.Watson [14] has combined traditional cryptography with DNA sequences to introduce hybrid security. The chains in a DNA have phosphate of one nucleotide and sugar of the next nucleotide to form a strand. DNA consists of two chains twisted around and form double strand helix. *A* and *T* are bond together while *C* and *G* bond in an opposite chain. Design information of DNA is transmitted to new cells during development and growth using complimentary pairing. The complimentary of the strands enables information to be replicated autonomously using a synthesizing template [1]. These complimentary strands are triplets of codons of nucleotides bases to form strands like figure 1;

A G G – C T C – A A G _T C C _ T A G

T C C _ C A G _ T T C _ A G G _ A T C

**Fig. 1. Complimentary DNA Strands**

DNA strands are mapped to numbers and alphabetical letters and other attributes and widely used for encoding and decoding as well as digital storing of data. Information encryption using DNA sequences can be used on the communication encryption methods, especially the ones in need of a robust data encryption scheme to challenge unauthorised access.

### *B. DNA translation, transcription*

When two DNA strands are separated by an enzyme, a single strand messenger RNA, complementary to DNA strand, is formed by mapping from DNA sequences, which consist of A, T, C, G, to complementary RNA sequences, which consist of U, A, G, C. These process non-coding segments, called introns in DNA sequences, are removed by splicing and remaining segments that encode information for protein synthesis, called exons, are assembled in mRNA [10].

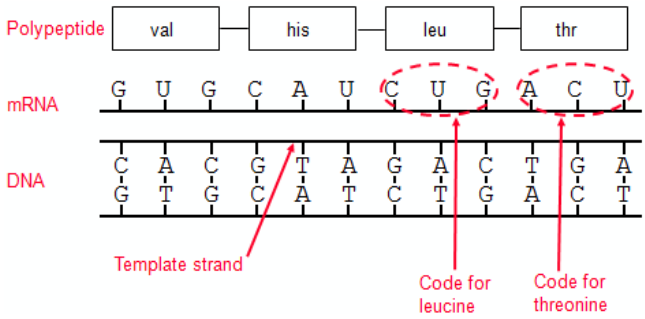


**Fig. 2. Translation Process from DNA to RNA& Protein [10]**

DNA and RNA both share common codons; A, G, C. DNA has an additional T codon whereas RNA has an additional U codon. Both these additional codons are used to form proteins. Figure 2 illustrates the simple concept of how three bases in DNA copied to mRNA by replacing T with U. The combination and sequence of the three letter codons of mRNA determines the order of the amino acids on the diagram. The figure below shows example of a protein construction through transcription and translation which converts to amino acids to construct a protein cell;

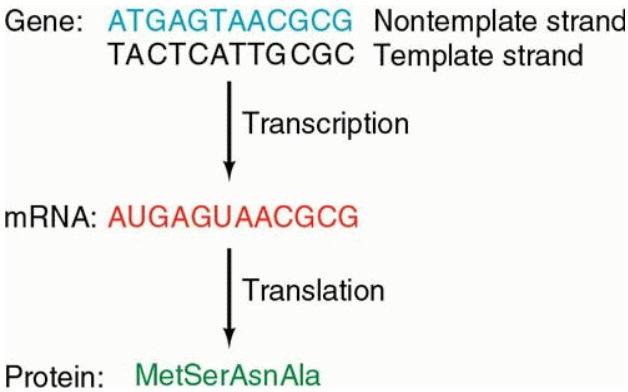


**Fig. 3. Transcription Process [10]**

Transcription is the synthesis of RNA from a DNA template as illustrated in figure 3. Only one strand of DNA is copied. A single gene may be transcribed thousands of times. After transcription, the DNA strands rejoin to form amino acids and subsequently protein.

### C. DNA Mutation:

When the sequence increases drastically, space complexity seems to be the major concern of dealing with a DNA searching system. Each of the DNA bases is converted to binary value before the matching process [9]. This enables to insert the one-time–pad or 'Exclusive-OR' algorithm in DNA sequences introduced by Gehani et al. If there is any error, a DNA strand is broken into segments and rearranges itself. In some cases the process involves deletion or insertion of some parts of DNA to form a correct codon base to recover from the error. These steps are taken to repair the damage to DNA, called mutation. DNA alignment is a fundamental comparison method to find common patterns between sequences, identify important regions, which consist of matching characters between two sequences or more, and positioning them correctly in a column. A count of the matching characters results in a measure of similarity between the sequences.[10]

## III. RELATED WORK

### A. Data Encryption Standard (DES)

DES is one of the modern cryptography algorithms to protect data. It is a symmetric encryption system which uses 64-bit blocks. The algorithm uses combination, substation and permutations between the text to be encoded and a key is generated. There is only 56 useful bits which they are used for key generation leading to $2^{56}$ different keys. Security of the DES algorithm relies on complexity of encryption key. With the powerful processors running on different machines they can find the possible key in a considerable amount of time [15].

First, the 128-bit key is partitioned into eight 16-bit sub blocks which are then directly used as the first eight key sub blocks. The 128-bit key is then intermittently shifted to the left by 25 positions, after which, the resulting 128-bit block is again partitioned into eight 16-bit sub-blocks to be directly used as the next eight key sub-blocks. The cyclic shift procedure described above is repeated until all of the required 52 16-bit key sub blocks have been generated[1] .

### B. Advanced Encryption Standard (AES

AES uses 128 bit block, a robust replacement for DES and supports larger size of blocks which is suitable to a wide range of systems, from smart cards with tiny memory capacities to large multiprocessor mainframe systems [16]. This type of algorithm supposed to protect the confidential but unclassified information against attacks, except brute-force attacks. Brute-force attack is process of systematic checking all possible key combinations until the correct key is found[11].

With the capability of new 128 bit AES key, it will take one billion years to break the key using a super computer, making AES secure against brute force attacks. However, public knowledge of the AES algorithm makes it vulnerable to known and unknown attacks. Practically this algorithm is not feasible to use on classified communications.

### C. Asymetric encryption with DNA technology

In DNA –Public Key Cryptography (PKC), there are two types of keys [12]; first one to be used for encryption and second one to create signature. Original message is encrypted using a public key and is decrypted by only those who own a private key. To create a signature a sender signs the message using a private key which is decrypted by only the corresponding public key. The keys in this method are biological molecules. The security is relied on the difficult biological problems. DNA PKC is immune from Quantum computer based attacks. Since it is impossible to replicate the cypher text, cloning can be prevented.

### D. One-Time-Pad using DNA

Gehani et all [4] introduced one-time-pads based on DNA to propose two encryption methods. One method uses substation technique to convert DNA sequences to cipher format using a pre-defined mapping table. The introduced key in one-time-pad is only used once making it impossible to break. The one-time-pad is stored in a large size library of codes, which is used once, and then discarded after that. Considering the limitation of this method, it is resilient to brute-force attacks and frequency analysis attacks, but as the library itself needs robust security, it is vulnerable to targeted attacks. If the database is accessed by unauthorised users, then the integrity and confidentiality of all communications can be compromised.

The algorithm for one-time-pad uses 'Exclusive-OR' operation of plain code and cipher key sequence. Exclusive-OR operation is impractical for DNA sequences. But Gehani et al. established one-time-pads by creating word pairs. All the algorithms for DNA encryption use pair-wise mapping to codon. It uses a substitution method using libraries of distinct pads, each of them which define specific pair-wise mapping keys. The method employs vast parallelism capability of DNA sequences and possible solutions represented by DNA strands.

### E. Primer

In this method a primer is designed to be transmitted between parties which represent block of texts. Sender and receiver create blocks of messages and choose a representation character, or words, for each block. Sharing the primer by both, the receiver will use the table to read the corresponding message. By applying the special function of primers to Polymerase Chain Reaction (PCR) amplification, the primers and coding mode are used as the key of the scheme [13]. Senders and receivers only communicate using primers, therefore the original message never transmitted over the public channel. This is secure because if the primer is found, the original message will not be obtained and revealed. Although this is a secure method to communicate sensitive information, several primers are required to fulfill communication demand. Also both parties should have access and agree on the original message and corresponding primer.

## IV. PPROPOSED ALGORITH

The proposed algorithm consists of several stages of conversions using different key generation tables as stated in figure 4. Detailed descriptions and representing tables included within this section;

Any information stored in a computer such as text and images, is in form of binary which represented by 0 and 1. Using the DNA and RNA alphabets explained in section 3, we

now can form a table to map these representations to binary bits as demonstrated in table 1.

**Table 1. Binary to DNA Conversion**

| S.No. | Bit1 | Bit 2 | DNA | RNA |
|---|---|---|---|---|
| 1 | 0 | 0 | A | A |
| 2 | 0 | 1 | C | C |
| 3 | 1 | 0 | G | G |
| 4 | 1 | 1 | T | U |

Data entry is converted to ASCII format. Then the ASCII format will be converted to DNA sequences, then RNA sequences using a key generated from corresponding look up tables of the sequences to be encrypted and sent over public channel. A secret key is generated using amino acid properties of the RNA sequences to be converted back to binary format cipher text.

Translating the 16 base RNA to amino acid is not possible as the amino acid sequences are three letters long. The example above shows 16 nucleotides which will make 5 amino acids with A, C, and U bases. We have remainder of U at the end of 15 letter long sequence. We can add complimentary letter like the one in chemical mechanism to insert messages to DNA strands. However, the reverse operation of the DNA to RNA conversion will reveal the strand, therefore exposing the message. Since every alphabet has representing RNA and DNA codons, therefore to fill the gap at the end of the string we assume representation of the codon letter as alphabet and insert 2 or 3 letter codon. If a sequence ended with T, then the 3 letter codon TTT will replace it. Since the insertion is done at the end of the string to make the sequence with 3 base, therefore the reverse operation will return the original message. This is exactly what happens in chemical processing of DNA replication when extra base pairs are inserted into a new place on the DNA sequences.

One can use this method to insert a message primer as explained in section III; this method also can be used to insert digital signature or digital certificate within a file or image to be sent over public channel.

After the DNA sequences is converted to RNA format, a secret key is generated from the chemical properties of the associated amino acids from table 3 and is used to convert the RNA format to binary sequences by substitution of the letters. The final Binary format is adding more complexity to the algorithm and more confusion to the cryptanalyst who will try to convert the binary to ASCII format. Since the DNA and RNA sequences as well as secret key generated from the amino acids are not easy to guess on the binary format. The cipher text then converted to binary format. The reverse operation is performed to decrypt the cipher text. Figure 4 illustrates the different operations and stages of conversion of messages to cipher format using associated tables. Although the final conversion of the cipher text to binary format will confuse any cryptanalyst, but the main strength of the proposed algorithm is the choice of the secret key from the amino acid properties of associated codons and sequences.

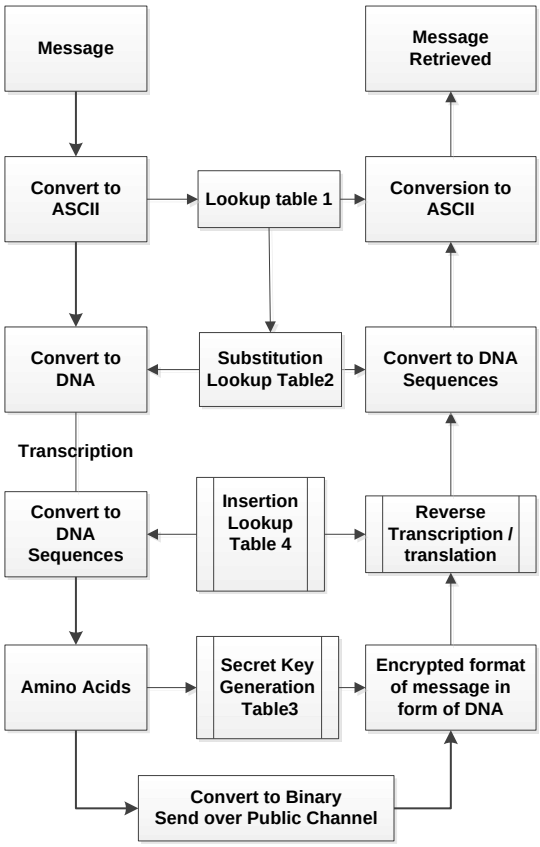


**Fig. 4. Encryption and Decryption Process**

*A. Look up Table and Key Genreration*

When DNA is transcribed, RNA is synthesized using this code. The RNA is a complimentary copy of one strand of the DNA. The RNA leaves the nucleus and in the cytoplasm it is translated into a protein. Each set of three contiguous RNA nucleotides codes for a single amino acid, and the protein is made of a chain of amino acids hooked to one another. Each set of three nucleotides in the DNA eventually codes for one amino acid in the final protein that is made from a given gene. The nucleotides and amino acids are not similar chemically, and it is the protein synthesis machinery of the cell that is needed to translate one code into the other. Table 2 is alteration of look up table from [1] which maps numeric and alphabetic values to codon triplets which is used as a look up table. Since this is widely used in almost all of the DNA-based cryptography [14] and easy to utilise by attackers, we use further secret key generation technique to extend the complexity. This technique will allocate a key from chemical properties of the DNA sequences and is obtained by both parties. If the message is decrypted and revealed by an attacker, the true meaning of the message will not be readable without this secret key. As the key choice and its method of use is decided on a prior agreement between two parties and never transmitted over public channel, it will be impossible to access. Other attributed and properties from the amino acids and chemical properties of the DNA strand can be used for data manipulation and substitution. This provides various choices for parties to agree on and use for data encoding, data encryption as well as data hiding.

**Table 2. Map Codon to Alphabets**

| Codon | English | Codon | English | Codon | English | Codon | English |
|---|---|---|---|---|---|---|---|
| AAA | a | CAA | q | GAA | G | TAA | W |
| AAC | b | CAC | r | GAC | H | TAC | X |
| AAG | c | CAC | s | GAG | I | TAG | Y |
| AAT | d | CAT | t | GAT | J | TAT | Z |
| ACA | e | CCA | u | GCA | K | TCA | 1 |
| ACC | f | CCC | v | GCC | L | TCC | 2 |
| ACG | g | CCG | w | GCG | M | TCG | 3 |
| ACT | h | CCT | x | GCT | N | TCT | 4 |
| AGA | i | CGA | y | GGA | O | TGA | 5 |
| AGC | j | CGC | z | GGC | P | TGC | 6 |
| AGG | k | CGG | A | GGG | Q | TGG | 7 |
| AGT | l | CGT | B | GGT | R | TGT | 8 |
| ATA | l | CTA | C | GTA | S | TTA | 9 |
| ATC | n | CTC | D | GTC | T | TTC | 0 |
| ATG | o | CTG | E | GTG | U | TTG | space |
| ATT | p | CTT | F | GTT | V | TTT | ,(period) |

There are 500 different types of amino acids. But there are only three bases to encode 20 amino acids [George Gamow biography] by living cells to build protein. A code of 3 nucleotides codes maximum of $4^3 = 64$ amino acids. The process of protein formation from DNA sequences is called transcription and translation as demonstrated on figure 3.

Table 3 demonstrates the information about biological properties of the chemical attributes of the DNA sequences which is used for public key generation.

**Table 3. Chemical Attributes of Amino Acids**

| Base | A | G | C | T | U |
|---|---|---|---|---|---|
| Name | Threonine | Alanine | Lysine | Valine | Glycine |
| Formula | $C_4H_9NO_3$ | $C_3H_7NO_2$ | $C_6H_{14}N_2O_2$ | $C_5H_{11}NO_2$ | $C_2H_5NO_2$ |
| 3 letter Symbol | Thr | Ala | Lys | Val | Gly |
| PH | 5.60 | 6.00 | 9.74 | 6.96 | 5.97 |
| H | 9 | 7 | 14 | 11 | 5 |
| N | 1 | 1 | 2 | 1 | 1 |
| O | 3 | 2 | 2 | 2 | 2 |
| C | 4 | 3 | 6 | 5 | 2 |
| MW | 119 | 89 | 146 | 117 | 75 |
| Weight | 119.12 | 89.10 | 146.19 | 117.15 | 75.07 |
| Residue Weight | 101.11 | 71.08 | 128.18 | 99.13 | 57.05 |
| Hydro-phobic value | 13 | 41 | -23 | 76 | 0 |
| Assigned English Letter | C | M | A | O | N |
| Code Triplet | ACU | GCU | AAA | GUU | GGU |

We propose chemical properties of the DNA sequences are used for key generation purposes which will be autonomously generated for communications. This will enhance the encryption security. There are many other properties and unique attributes of the common amino acids as listed in table 3, which can be used as indexing table to generate a secret key. This process relies on agreement by both parties in advance on the choice, which provides numerous choices of keys in order to be used for information encoding. This technique can replace many key generation mechanisms used for digital signatures, one-time-pad techniques, and any other techniques which employ DNA sequences for encryption and information hiding purposes.

**Table 4. Codon Mapping to Amino Acids**

| DNA Codon | T | C | A | G |
|---|---|---|---|---|
| T | TTT, TTC } Phe<br>TTA, TTG } Leu | TCT, TCC, TCA, TCG } Ser | TAT, TAC } The<br>TAA, TAG } STOP | TGT, TGC } Cys<br>TGA } STOP<br>TGG } Trp |
| C | CTT, CTC, CTA, CTG } Leu | CCT, CCC, CCA, CCG } Pro | CAT, CAC } His<br>CAA, CAG } Gln | CGT, CGC, CGA, CGG } Arg |
| A | AAT, ATC, ATA } IlE<br>ATG } Met | ACT, ACC, ACA, ACG } Thr | AAT, AAC } Asn<br>AAA, AAG } Lys | AGT, AGC } Ser<br>AGA, AGG } Arg |
| G | GTT, GTC, GTA, GTG } Val | GCT, GCC, GCA, GCG } Ala | GAT, GAC } Asp<br>GAA, GAG } Glu | GGT, GGC, GGA, GGG } Gly |

Combining the chemical attributes of the amino acids in table 3, alphabetical representation of codons listed in table 2 and representing amino acids in table 4 can have massive capacity for data manipulation, substitution, and insertion to extended key generation which can be used to encode confidential information sent over public channel.

## V. DISCUSSION

The traditional library of codebooks consists of large number of one-time-pads. This in turn contributes to a long portion substitution in converting the plain text to DNA sequences making this method of encryption unbreakable, but at the same time inefficient. With an eye on the well-known principle "*the security of the crypto scheme is in key management, not secrecy of the algorithm."* (Kahn, The code breakers) it is imperative to sensibly manage key and carefully choose a secret key rather than design a robust and complicated algorithm. In the proposed algorithm a message is encrypted through several stages of conversion to different formats with insertion of secret key from a look up table. This will add more complexity and complication to encryption mechanism. The one-time-pad algorithm prevents future penetration and is resilient to frequency analysis based attacks, the limitation of choice is the real constrain for the algorithm. Numerous communications requires several one-time-pads. However use of random choice from a table which is generated based on the plain text content will be more flexible, efficient and furthermore, hard to access by cryptanalysts that use frequency based attacks.

## VI. CONCLUSION

Most of the known algorithms for DNA encryption use secret codes from one-time-pads generated from DNA sequences as a secret key to encrypt data. We use chemical properties of the DNA sequences of the cipher text to encrypt

data over public channel to add key extension and complexity to the encryption algorithm. This technique will add uncertainty to the key and message exploitation. Furthermore, if the cipher text is accessed and content is revealed, the true meaning of the message will not be revealed without the key. This technique will enhance the security of the encryptions and encryption algorithms.

## VII. Future Work

For the future work we will use experimental analysis of the proposed algorithm to investigate alternative properties of the chemical biology of the sequences used for key generation. Also a thorough analysis on complexity of the algorithm and added overhead will be carried out comparing the objectives to existing algorithms.

## VIII. References


[1] P. Rakheja, "Integrating DNA Computing in International Data Encryption Algorithm ( IDEA )," vol. 26, no. 3, pp. 1–6, 2011.

[2] Behr, F., Fossum, V., & Mitzenmacher, M, "Estimating and Comparing Entropy across Written Natural Languages Using PPM Compression", Technical Report TR-12-02, Harvard University, 2002

[3] L.M. Adleman, Molecular Computation of solutions to combinatorial problems, '*Science*', 266:1021-1024, November 1994

[4] A. Gehani, T. Labean, and J. Reif, "DNA-Based Cryptography ∗," pp. 1–17, 2000.

[5] S. V Kartalopoulos, "DNA-INSPIRED CRYPTOGRAPHIC METHOD IN OPTICAL COMMUNICATIONS", 2008

[6] D. Heider and A. Barnekow, "DNA-based watermarks using the DNA-Crypt algorithm.," *BMC Bioinformatics*, vol. 8, p. 176, Jan. 2007.

[7] J. D. Watson, F. H. C. Crick, "A structure for de oxy ribose nucleic acid", Nature, Vol. 25, 99. 737-738, 1953

[8] A. Azfar, K.-K. R. Choo, and L. Liu, "A Study of Ten Popular Android Mobile VoIP Applications: Are the Communications Encrypted?," *2014 47th Hawaii Int. Conf. Syst. Sci.*, pp. 4858–4867, Jan. 2014.

[9] C. Kreibich and J. Crowcroft, "Efficient sequence alignment of network traffic," *Proc. 6th ACM SIGCOMM Internet Meas. - IMC '06*, p. 307, 2006.

[10] S. Lloyd and Q. O. Snell, "Sequence Alignment with Traceback on Reconfigurable Hardware," *2008 Int. Conf. Reconfigurable Comput. FPGAs*, pp. 259–264, Dec. 2008.

[11] Mohit Arora, "How secure is AES against brute force attacks?", Free scale Semiconductor, EE Times, 2012

[12] LAI XueJia, LU MingXin "Asymmetric encryption and signature method with DNA technology" Vol. 53 No.3:506-514 March 2010

[13] G. Cui, L. Qin, Y. Wang, and X. Zhang, "An encryption scheme using DNA technology," *2008 3rd Int. Conf. Bio-Inspired Comput. Theor. Appl.*, pp. 37–42, Sep. 2008.

[14] A. Cherian, S. R. Raj, and A. Abraham, "A Survey on different DNA Cryptographic Methods," vol. 2, no. 4, pp. 167–169, 2013.

[15] DES Encryption, www.kioskeas.net, 2014

[16] MDSN Magazine, Microsoft corporation publication 2003