International Journal of Bioinformatics and Biomedical Engineering
Vol. 2, No. 3, 2016, pp. 46-50
http://www.aiscience.org/journal/ijbbe
ISSN: 2381-7399 (Print); ISSN: 2381-7402 (Online)

Protein Sequence Safe Transfer Using Graph Operation

M. Yamuna*, K. Karthika

School of Advanced Sciences, Vellore Institute of Technology, Vellore, India

Abstract

Any change in genetic instruction leads to malfunctioning of mRNA. Malfunctioning of mRNA induces various diseases like mutation identification of these usually done using DNA samples. In such situation for demand of genetic privacy safe transfer of the data is mandatory. Graph theory has applications in various fields, there are numerous property that a graph satisfies. In this paper we provide a new technique to safe transfer, details of a protein sequence as a weighted graph using graph operation and graph domination.

Keywords

Genetic, Protein, DNA, Graphs

Received: September 11, 2015 / Accepted: October 27, 2015 / Published online: June 17, 2016

1. Introduction

Encryption is the process of translating plain text data (plaintext) into something that appears to be random and meaningless (ciphertext). Decryption is the process of converting ciphertext back to plaintext. To encrypt more than a small amount of data, symmetric encryption is used. A symmetric key is used during both the encryption and decryption processes. To decrypt a particular piece of ciphertext, the key that was used to encrypt the data must be used. The goal of every encryption algorithm is to make it as difficult as possible to decrypt the generated ciphertext without using the key. [1].

Graph theory in cryptography have recently gained momentum. In [2], Monika Polak et al have provided the algebraic constructions of regular graphs of large girth and graphs with large cycle indicator. They have described some algorithms of coding theory and cryptography based on such special families of graphs. In [3], Wael Mahmoud Al Etaiwi has provided an encryption algorithm to encrypt and decrypt data securely with the benefits of graph theory properties. M. Yamuna et al have proposed the method for encrypting any binary string using cipher chain blocking. Also they have used a musical note as key to construct the degree sequence of the graph [4].

In near future genetic advances will play an important role in our lives. But this advance will raise privacy challenges. It is true that genetic information promises to enhance the qualities of life better understand ourselves improves healthcare that can benefit individuals, families and in general our society. But genetic information can also be misused in many ways, which may lead to embarrassment or distressed due to use or disclosure of genetic information. So, preserving details regarding these data are important but challenging.

In [5], B. Claerhout et al have provided some of the privacy – protection problems related to classical and genomic medicine, and highlights the relevance of trusted third parties and of privacy – enhancing techniques in the context of data collection. In [6], provides details of the impact of the personal data production act 2010. It also considers, whether the various personal data protection principles are applicable to the act of DNA profiling and the creation of bioinformatics. In [7], Clementine Gritti et al have introduced

* Corresponding author
E-mail address: myamuna@vit.ac.in (M. Yamuna), karthika.k@vit.ac.in (K. Karthika)

a solution to encrypt data using the DNA sequences of the sender and to decrypt data using the DNA sequences of the receiver. Mete Akgun et al have done a survey on privacy preserving processing of genomic data [8]. In this paper we provide a method of encrypting a protein sequence as a weighted graph, which can be send in pubic domain.

2. Preliminary Note

In this section we provide the basic results of DNA, protein sequence and graph theory which are required for proposed encryption scheme.

2.1. DNA

Deoxyribonucleic acid, a nucleic acid that consists of two long chains of nucleotides twisted together into a double helix and joined by hydrogen bonds between complementary bases adenine and thymine or cytosine and guanine; it carries the cell's genetic information and hereditary characteristics via its nucleotides and their sequence and is capable of self – replication and RNA synthesis. [9]

2.2. Protein Sequence

Peptide sequence, or amino acid sequence, is the order in which amino acid residues, connected by peptide bonds, lie in the chain in peptides and proteins. The sequence is generally reported from the N-terminal end containing free amino group to the C-terminal end containing free carboxyl group. Peptide sequence is often called protein sequence if it represents the primary structure of a protein. [10].

2.3. Graph

In the most common sense of the term, a graph is an ordered pair G = (V, E) comprising a set V of vertices or nodes together with a set E of edges or links, which are 2 – element subsets of V (that is an edge is related with two vertices, and the relation is represented as an unordered pair of the vertices with respect to the particular edge).

2.4. Weighted Graph

A graph is a weighted graph if a number (weight) is assigned to each edge. Such weights might represent, for example, costs, lengths or capacities, etc. depending on the problem at hand. Some authors call such a graph a network [11].

2.5. Dominating Set

A set of vertices D in a graph G = (V, E) is a dominating set if every vertex of V – D is adjacent to some vertex of D. If D has the smallest possible cardinality of any dominating set of G, then D is called a minimum dominating set – abbreviated MDS. The cardinality of any MDS for G is called the domination number of G and it is denoted by γ (G). γ - set denotes a dominating set for G with minimum cardinality.

2.6. Edge Contraction

For a pair of vertices u, v of G, denote by G.uv the graph obtained by identifying u and v. Let (uv) denote the identified vertex. So G.uv may be viewed as the graph obtained from G by deleting the vertices u and v and appending a new vertex, denoted by (uv), that is adjacent to all the vertices of G – u – v that were originally adjacent to either of u or v [12]. The resulting graph if a multiple graph is retained back as a simple graph.

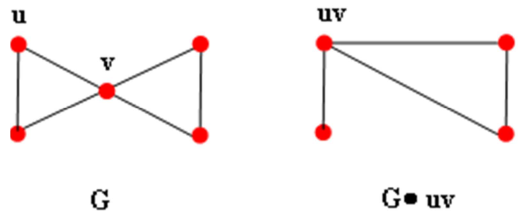

Figure. 1. Example of edge contraction.

In Fig. 1, we remove the edge (u v) and its two incident vertices, u and v are merged into a new vertex uv. The edges incident to u and v in G, is now incident on uv in G.uv.

Graph operations are extensively used in various research domains. In this paper we choose graph operations which satisfy property that the domination number of a graph does not change on edge contraction. We have graphs where every edge in the graph satisfies the property, some edges satisfy the property and no edge satisfies. If every edge in the graph satisfies the property, then the graph is said to domination dot stable graphs [13].

In all Figures

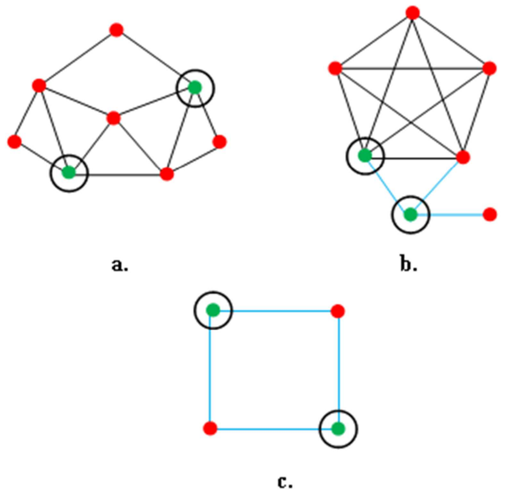

Figure. 2. Example of no / some / all edges are contractible.

🟢 - Represent a vertex that belongs to D.
🔴 - Represent a vertex that belongs to V – D.

In Fig. 2, in all three graphs, black color edges represent edges such that the domination number does not change on edge contraction.

We use graphs having atleast one edge with this property for safe transfer of genetic privacy details which is available as a DNA or protein sequence.

3. Materials and Methods

Proposed Encryption Method

In this section we provide a method of encryption and decryption of a protein sequence as a weighted graph.

3.1. Edge Weight Assignment

Modular arithmetic can be handled mathematically by introducing a congruence relation on the integers that is compatible with the operations of the ring of integers: addition, subtraction, and multiplication. For a fixed modulus n, it is defined as follows.

Two integers a and b are said to be congruent modulo n, if their difference a − b is an integer multiple of n. If this is the case, it is expressed as:

$$a \equiv b \bmod (m)$$

The above statement is read: "a is congruent to b modulo n".

In any DNA sequence, there are only four bases A, T, G, C. By the above definition, under addition modulo 4, the possible remainders are 0, 1, 2, 3. So the numbers having these values as remainders can be used for these bases. We use the following for the bases that is, we can use any number whose remainder value is 1 when divided by four in the place of A, any number whose remainder value is 2 when divided by four in the place of T, any number whose remainder value is 3 when divided by four in the place of G, any number whose remainder value is 0 when divided by four in the place of C respectively.

$A \equiv 1 \pmod 4$, ($A = \{1, 5, 9, 13, 17, 21, 25, 29, 33, \ldots\}$).

$T \equiv 2 \pmod 4$, ($T = \{2, 6, 10, 14, 18, 22,$

$26, 30, 34, \ldots\}$).

$G \equiv 3 \pmod 4$, ($G = \{3, 7, 11, 15, 19, 23, 27, 31, 35, \ldots\}$).

$C \equiv 4 \pmod 4$) , ($C = \{4, 8, 12, 16, 20, 24, 28, 32, 36, \ldots\}$).

3.2. DNA Codon Table

The genetic code is traditionally represented as an RNA codon table because, when proteins are made in a cell by ribosomes, it is mRNA that directs protein synthesis [14]. Table – 1 provides the standard DNA Codon table which we shall use for our proposed method.

Table 1. Standard Genetic Code.

	T		C		A		G		
T	TTT	F	TCT	S	TAT	Y	TGT	C	T
	TTC		TCC		TAC		TGC		C
	TTA	L	TCA		TAA	Stop	TGA	Stop	A
	TTG		TCG		TAG		TGG	W	G
C	CTT	L	CCT	P	CAT	H	CGT	R	T
	CTC		CCC		CAC		CGC		C
	CTA		CCA		CAA	Q	CGA		A
	CTG		CCG		CAG		CGG		G
A	ATT	I	ACT	T	AAT	N	AGT	S	T
	ATC		ACC		AAC		AGC		C
	ATA		ACA		AAA	K	AGA	R	A
	ATG	M	ACG		AAG		AGG		G
G	GTT	V	GCT	A	GAT	D	GGT	G	T
	GTC		GCC		GAC		GGC		C
	GTA		GCA		GAA	E	GGA		A
	GTG		GCG		GAG		GGG		G

3.3. Encryption Algorithm

Let P be the sequence to be encrypted. Consider the example, P: M H G N L be the sequence to be encrypted.

Step 1 Convert the protein sequence P into DNA sequence P1 using Table – 1.

For sequence P, P1: ATGCACGGCAATTTG

Step 2 Replace each base in the sequence P1 by any random value from Sec. 3.1, taking care that the numbers are in the increasing order, to generate a sequence P2. P2 is a sequence of numbers. Let length of P2 = k = $\{w_1, w_2, \ldots, w_k\}$. Note that $w_1 < w_2 < \ldots < w_k$.

For sequence P, P2 = 1 2 3 4 5 8 11 15 16 17 21 22 26 30 35. Length of P2 = 15

Step 3 Choose a graph G with atleast k edges satisfying the condition $\gamma(G) = \gamma(G_\bullet uv)$.

For sequence P we choose the random graph in Fig. 3.

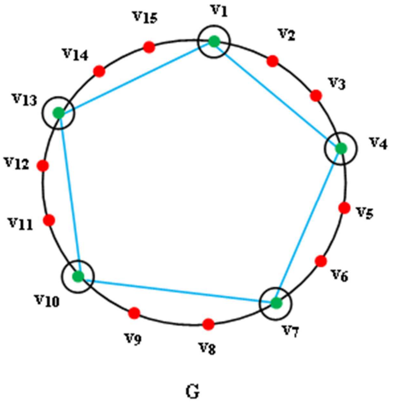

Figure 3. Random graph with 20 edges.

In this graph we have 15 $\{(v_1\ v_2), (v_2\ v_3), (v_3\ v_4), (v_4\ v_5), (v_5\ v_6), (v_6\ v_7), (v_7\ v_8), (v_8\ v_9), (v_9\ v_{10}), (v_{10}\ v_{11}), (v_{11}\ v_{12}), (v_{12}\ v_{13}), (v_{13}\ v_{14}), (v_{14}\ v_{15}), (v_{15}\ v_1)\}$edges satisfying the condition $\gamma(G) = \gamma(G_\bullet uv)$.

Step 4 Assign the weights w_1, w_2, …, w_k to the stable edges in the graph G randomly to generate a new graph G_1. For sequence P the weights are assigned arbitrarily as seen in Fig. 4.

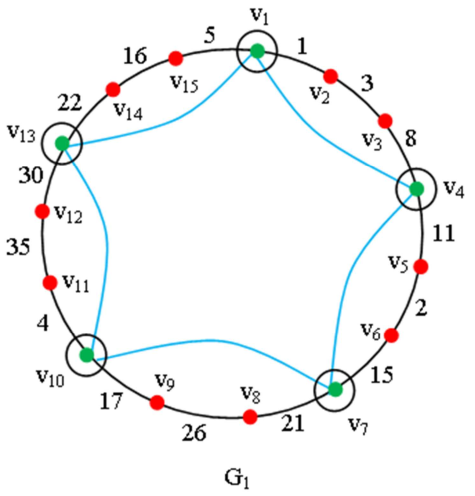

Figure 4. Graph with weights assigned to contractible edges.

Step 5 Assign arbitrary weights for the remaining edges if any in the graph G_1 to generate a weighted graph G_2.

For our sequence P, the arbitrary weights are assigned to blue edges as seen in Figure. 5.

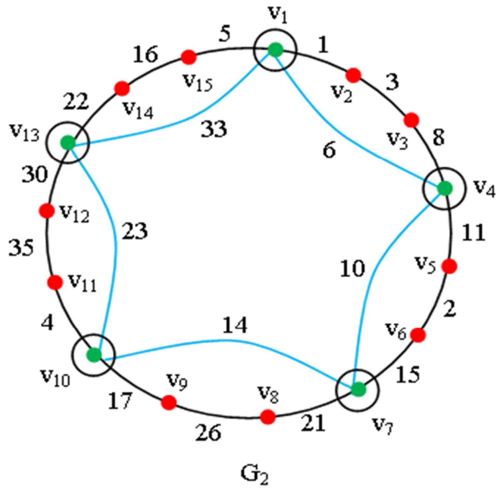

Figure 5. Graph to be encrypted.

Step 6 Send the graph G_2 to the receiver.

3.4. Decryption Algorithm

By reversing the procedure we can decrypt the graph into its corresponding protein sequence.

For example let the received G_3 be graph as seen in Fig. 6. Here $\gamma(G) = 2$.

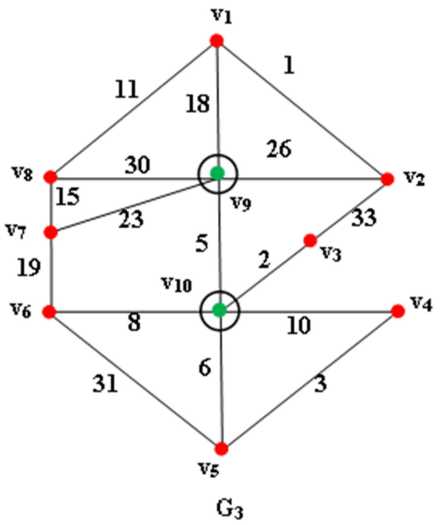

Figure 6. Received graph to be decrypted.

Step 1 List out the edge weights of the stable edges in increasing order. The stable edges are $\{(v_1\ v_2), (v_1\ v_8), (v_1\ v_9), (v_2\ v_3), (v_2\ v_9), (v_3\ v_{10}), (v_4\ v_5), (v_4\ v_{10}), (v_5\ v_6), (v_5\ v_{10}), (v_6\ v_7), (v_6\ v_{10}), (v_7\ v_8), (v_7\ v_9), (v_8\ v_9)\}$. The weights corresponding to these edges are {1, 11, 18, 33, 26, 2, 3, 10, 31, 6, 19, 8, 15, 23, 30}. Arranging them in increasing order we generate

P2: 1 2 3 6 8 10 11 15 18 19 23 26 30 31 33.

Step 2 Convert P2 into a DNA sequence as discussed in Sec. 3.1.

The resulting DNA sequence for P2 is

P1: ATGTCTGGTGGTTGA

Step 3 Convert P1 into protein sequence by using Table – 1.

The corresponding protein sequence is P: M S G G

4. Conclusion

Protein sequences can be of any length. They represent the genetic details and hence interrelated with the genetic privacy. Safe transmission of these details is mandatory in most cases. A graph has a special property that it can be of any size and include any number of edges. In our proposed method we have used edges for protein encryption.

A protein has twenty amino acids while a DNA has only four nucleotide base A, T, G, C. In modulo 4, addition and multiplication the possible remainders are 0, 1, 2, 3. We use this remainder property as edge values. So, a protein sequence is represented as a graph.

i. Not all the edges are used. Only those edges whose domination number does not change by contradiction are used. This improves the security of the transmitted sequence.

ii. A graph can have maximum number of $\frac{n\,(n-1)}{2}$ edges. So, to encrypt sequence of length 1225, we need a graph with only 50 vertices.

iii. Weighted graphs are used for various reasons in multidimensional fields. Hence numerous weighted graphs are available in public domain for various reasons. This enables safe transfer of the genetic details.

iv. Suppose, if we consider a graph with 10 edges, then we can arrange the edge weights into 10! ways. So, the sequence can be encrypted into 3628800 ways. This means that decryption becomes more difficult.

So we conclude that the proposed method is good for safe transmission of any protein sequence.

References

[1] https://msdn.microsoft.com/en-us/library/windows/desktop/aa381939%28v=vs.85%29.aspx.

[2] Monika Polak, Urszula Romanczuk, Vasyl Ustimenko, Aneta Wroblewska. (2013). On the Applications of Extremal Graph Theory to Coding Theory and Cryptography, *Electronic Notes in Discrete Mathematics* Vol. 43, pp. 329–342.

[3] Wael Mahmoud Al Etaiwi. (2014). Encryption Algorithm Using Graph Theory, *Journal of Scientific Research & Reports*, 3(19), pp. 2519–2527.

[4] Yamuna, M. Sankar, A. Siddarth Ravichandran, Harish V. (2013). Encryption of a Binary String Using Music Notes and Graph theory, *International Journal of Engineering & Technology*, Vol. 5(3), pp. 2920–2925.

[5] Claerhout, B. DeMoor, G. J. E. (2005). Privacy Protection for Clinical and Genomic Data, the Use of Privacy – Enhancing Techniques in Medicine, *International Journal of Medical Informatics,* 74, pp. 257–265.

[6] Ida Madieha Azmi. (2011). Bioinformatics and Genetic Privacy: The Impact of the Personal Data Protection Act 2010, *Computer law & security review*, 27, pp. 394–401.

[7] Clementine Gritti, Willy Susilo, Thomas Plantard, Khin Than Win. (2015). Privacy – Preserving Encryption Scheme Using DNA Parentage Test, Theoretical Computer Science, 580, pp. 1–13.

[8] Mete Akgun, Osman Bayrak, A. Bugra Ozer, Samil Sairoglu. (2015). Privacy Preserving Processing of Genomic Data: A Survey, Journal of Biomedical Informatics, 56, pp. 103–111.

[9] http://medical-dictionary.thefree dictionary.com/DNA.

[10] http://encyclopedia.thefreedictionary.com/Protein+sequence.

[11] http://en.wikipedia.org/wiki/Graph_%28mathematics%29.

[12] Burton, T. Sumner, D. (2006). Domination Dot Critical Graphs, *Discrete Math.* 306, pp. 11-18.

[13] Yamuna, M. Karthika, K. (2011). Excellent – Domination Dot Stable Graphs, *International Journal of Engineering Science, Advanced Computing and Bio – Technology*, 2(4), pp. 209-216.

[14] https://en.wikipedia.Org/wiki/DNA_codon_table.