A Novel Technique to Hide and Encrypt The Secret Information in DNA

Basim Sahar Yaseen
Department of computer sciences ,Satt alarab university college
basim_yaseen2000@yahoo.com

Abstract

The paper suggests a mathematical manner that simulates the transition processes of the hereditary information from/to its storage in the cell nucleus (genetic bases in DNA strand), and via central dogma processes. The suggested framework begins with a real mRNA codons sequence to recode the secret message from alphabet to real number values by using their statistics ,and then ,these real values will be stored in the genetic bases bonds in a novel technique and, via this step we can construct a numerical DNA strand for the original secret message ,that it can be manufactured biologically in the laboratory .The aim of the proposal technique is finding the mathematical framework which utilizes the variation fundamental of bonds power values between genetic bases to hide and to encrypt the secret information, then reproducing the DNA strand of the mRNA with the message. In this paper ,we describe the design of a DNA coder mechanism that simulates the processes of central dogma inverse operation to hide and to encrypt an information via this operation.

Keywords : DNA , mRNA , statistics ,genetic bonds ,central dogma .

الخلاصة

يقترح البحث اسلوب رياضي يحاكي معالجات النقل للمعلومات الوراثية من / والى اماكن خزنها في نواة الخلية الحية (القواعد الجينية في سلسلة DNA) ,ومن خلال معالجات العقيدة المركزية داخل الخلية.يبدأ الهيكل الرياضي المقترح بسلسلة حقيقية من ترميزات mRNA لاعادة ترميز الرسالة السرية ونقلها من الابجدية الى قيم رقمية حقيقية, تخزن هذه القيم بالروابط بين القواعد الجينية باستخدام احصاءاتها ,ثم تقوم هذه الخطوة بانتاج سلسله رقمية من DNA تعبر عن الرسالة السرية الاصلية, والشكل الجديد للرسالة يمكن ان ينتج مختبريا كسلسلة حياتية ., الهدف من التقنية المقترحه هو ايجاد هيكل رياضي يخدم التنوع في قيم قوة الربط بين القواعد الجينية وتوظيفها في اخفاء وتشفير المعلومات السرية ,وكذلك اعادة انتاج DNA للـ mRNA الخاصة بالرساله .وفي هذا البحث قمنا بتوضيح كيفية تصميم مرمز DNA يحاكي معالجات العقيدة المركزية المعكوسة من سلسلة الاحماض الامينية الى سلسلة DNA الحاوية على المعلومات السرية.

الكلمات المفتاحية : mRNA, DNA, الاحصاءات , الروابط الجينية , العقيدة المركزية

1.Introduction

1.1 Data hiding and Encryption by using DNA

With the birth of molecular biology 70 years ago came the realization that the processes within biological cells are carried out by molecular machines ,and that the most central processes involved the manipulation of information-bearing polymers. The term DATA HIDING(Hayam and etal 2011, Samir and etal 2011) defines distinctly prevailing manners that can be used to hide the secret data in a digital media ,whether a text , image , or any form of the data .Occasionally , the media is not classic, for instance DNA. Whereas ,the term DATA ENCRYPTION (Bruce 1997 ,Henk 2007) refers any mathematical steps disarrange plain text characters ,instead of hide it in a cover ,on the strength of potency of the hiding ,or encrypting we have to choose the technique ,and work tools, sometimes, security applications will have been combining the two concepts types . From a more global perspective , however , it can be stated that the fundamental reason behind combined protecting is to be better able to cope with the large-scale attacking threats(Bruce 1997) ,that we face today, by using a variation of the well-known encrypt-and-hide rule. So, when we survey all works which use genetics in the information protection, we can abbreviation summarize features of successful algorithm by:

1- Basically, work should be depend on genuine hereditary information (genetics).
2- The suggested algorithm should have a maximum potency to be suitable for the security of the information, as well as , it keeps the genetics safety .
3- To keep up with the simplicity and perfect evaluations ,in the two ways encrypt/hide and decrypt/get back information.

In addition to the previous features of encryption and data hiding should be achieved. Essentially, we can classify all genetic works ,in security, as:

1- Some of them, restricted to digital work merely.
2- The others, adopts the bio-digital works.

1.2 Biological Introduction

The heredity information in the nucleic can be viewed as strands contains a large amount of connected genetic bases called DNA that are coding the genetic information, the DNA molecule is a polymer composed of nucleotides. Each of these nucleotides is a letter in the "genetic alphabet". Each nucleotide is formed of a phosphate group, a sugar and a nitrogenous base . Two nucleotides are connected by a phosphodiester bond, in which a phosphate group links the 3'-hydroxyl group of one nucleotide to the 5'-hydroxyl group of the next, giving rise to directionality in the polynucleotide chain (Paolella 1998, Watson and etal 1953) .The base defines as a type of nucleotide. There are two types of bases in DNA: purines and pyrimidines. The two main purines that occur in DNA are adenine (A) and guanine (G). As for pyrimidines, cytosine (C) and thymine (T) are typically found . Nitrogeneous bases occurring in DNA: adenine (A) and guanine (G) are purines, whereas cytosine (C) and thymine (T) are pyrimidines[5,6].The bonds between genetic bases can be exploited to encode the hiding digital information ,just as in the present work , by using a series of processes called central dogma (Paolella 1998, Watson and etal 1953), the DNA arrangement translates to a new forms , that is a sequence of mRNA codons . mRNA sequence divided into triple codons ,each one of them is interrelated with suitable amino acid via a genetic table (Watson and etal 1953),it has 64 codons and 20 amino acids. The bonds simulation mechanism at present proposal is finding a mathematical framework to encode message numbers for encryption, coding,or hiding purposes ,this mechanism will construct a map of bonds values and get yields can be considered a DNA strand organization .Realization of the novelty property is via the previous characteristics of the proposal mechanism.

2 Related works

In hiding data by using mutates (Prased and etal 2011),this way benefit from alteration operation of the last codon in the mRNA sequence that amino acid stem from it, and the alteration to keep up with subversion of the original amino acid sequence, this manner used in codons of the genetic table, and it can be considered as a simple. One more way, character statistics (Feng and etal 2002) ,this way depend on manner nominates for each character a codon from genetic table via the frequencies of letters and codons, and then, it will encrypt the the message by using created table, this way can be considered as a simple and powerless. Other manner (Motamemi and etal 2007) ,the rule is searching in codons sequence for codon has a form(BAB˜), where B˜ is a complement base of B, and depending on some specific rules that act the hiding process , this manner can be

considered more difficulty from previous first two manners, but it's the best and powerful from them .

3 The proposal technique

We can draw a summary sketch of the suggested technique stages, figure no.1 ,shows the stages of the mathematical technique, as follows :

Note: the mRNA sequence should be a real cell information (laboratory information).

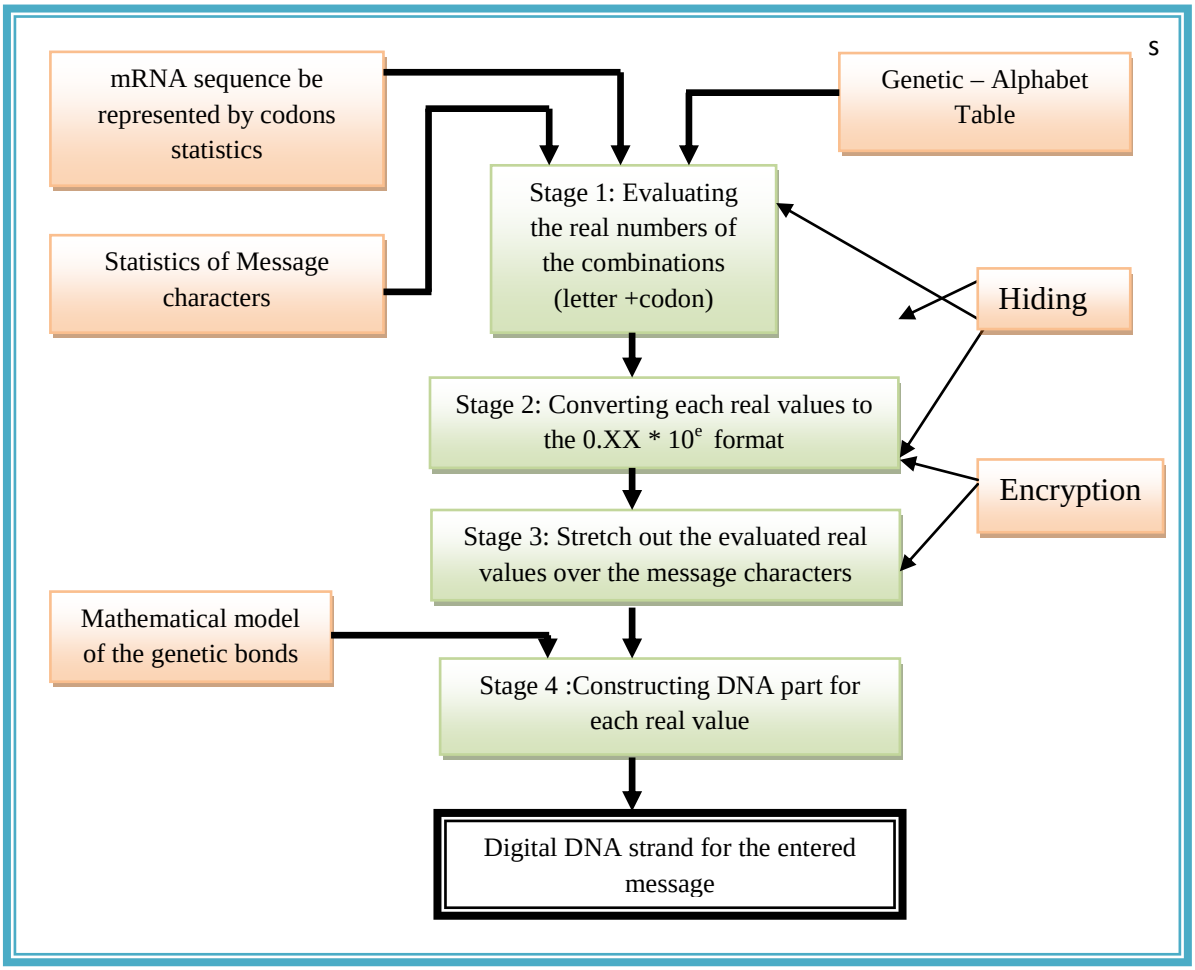

Figure 1 : General outline for the suggested technique stages

3.1 How to construct the Genetic – Alphabet Table

Table 1 : The Genetic – Alphabet Table

G		A		C		U		
UGU Cys[C]	A	UAU Tyr[Y]	P	UCU Ser[S1]	3	UUU Phe[F]	:	U
UGC Cys[C]	B	UAC Tyr[Y]	Q	UCC Ser[S2]	4	UUC Phe[F]	~	
UGA Stop[O]	notuse	UAA Stop[O]	notuse	UCA Ser[S3]	5	UUA Leu[L4]	!	
UGG Try[W]	C	UAG Stop[O]	notuse	UCG Ser[S4]	6	UUG Leu[L5]	@	
CGU Arg[R1]	D	CAU His[H]	R	CCU Pro[P1]	7	CUU Leu[L1]	#	C
CGC Arg[R2]	E	CAC His[H]	S	CCC Pro[P2]	8	CUC Leu[L2]	$	
CGA Arg[R3]	F	CAA Gln[Q]	T	CCA Pro[P3]	9	CUA Leu[L3]	%	
CGG Arg[R4]	G	CAG Gln[Q]	U	CCG Pro[P4]	,	CUG Leu[L4]	^	
AGU Ser[S5]	H	AAU Asn[N]	V	ACU Thr[T1]	.	AUU IIe[I]	&	A
AGC Ser[S6]	I	AAC Asn[N]	W	ACC Thr[T2]	<	AUC IIe[I]	*	
AGA Arg[R5]	J	AAA Lys[K]	X	AGA Thr[T3]	>	AUA IIe[I]	(	
AGG Arg[R5]	K	AAG Lys[K]	Y	ACG Thr[T4]	/	AUG Met[M]	notuse	
GGU Gly[G1]	L	GAU Asp[D]	Z	GCU Ala[A1]	?	GUU Val[V1]	)	G
GGC Gly[G2]	M	GAC Asp[D]	0	GCC Ala[A2]	“	GUC Val[V2]	[	
GGA Gly[G3]	N	GAA Glu[E]	1	GCA Ala[A3]	}	GUA Val[V3]	]	
GGG Gly[G4]	O	GAG Glu[E]	2	GCG Ala[A4]	;	GUG Val[V4]	{	

As well known (and as in 1.2 section), the genetic table is consisting of 64 codons from genetic bases (A,U,C,G),these codons distributed on 20 amino acid.So, we can disperse the letters of the used alphabet (English) over the codons in this table ,except the begin codon(AUG) and the stop codons(UGA,UAA,UAG),the resulted table will use to supplement the codons sequence ,so long as insufficient chain.

When message has been read ,the table of letters statistics created ,however, a digital mRNA can be prepared from biological laboratory ,and its codons frequency intended ,and when we opposite two statistics tables , two possibilities will appear ,that are:

1- Number of letters in the message ,will be equivalent to number of codons in the mRNA sequence ,in this case ,we will transfer control to the next stage ,without taking advantage of genetic-alphabet table.
2- Number of codons in the mRNA sequence less than the number of letters in the message ,at that time, we will take advantage of this table.

3.2 Building the mathematical model of the Genetic Bonds

As well known, the genetic bonds join the genetic bases together, and it joins the gens in the chromosome too. Mathematically and novelty, we are building a mathematical model for these bonds to store and to encode the secret information(message) that it represented by real number series, this model simulates the real bonds in the biological strands, and the figure no. 2 illustrates that.

The shape of the proposal bond between genetic bases and the distributed values can be viewed as shown in the figure 3 :

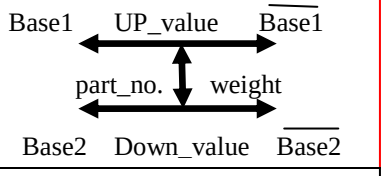

Figure 3:proposal bond

3.3 Stages of the technique

Stage no.1 is Evaluating the real numbers of the combinations(letter-codon),the statistics of the message letters are to be arranged descending, and by same style codons statistics are to be arranged too, for each letter arrangement is dividing over codon frequency that corresponded to it.

$R\_No[i] = L\_arrangement[i] / Cod\_freq[i]$ ….. (1)

In the stage 2 ,R_No series is converting to $0.xx * 10^{e}$ format ,and then ,the resulted numbers series will stretch out on the all message letters ,via stage 3 procedure ,the last stage (stage 4) is construction DNA strand for real numbers series as follows :

For each real number will build a part from the final DNA strand , this part consists of two portions : first is a head ,that it can be selected from convenient mathematical model areas ,the head part represents the power of ten, it has long one gene, and it can calculate from:

From selected gene, $e = Po_{up} - Po_{down}$ …. (2)

And, $Y_i = rr$ …. (3) , where Y_i is the initial value of 0.xx number part, and rr is the real value of the gene. And the other is tail ,it can be computed from one or more genes via the equation:

$Y_i = Y_i + (Po_{up} + Po_{down}) * rr.$ …. (4)

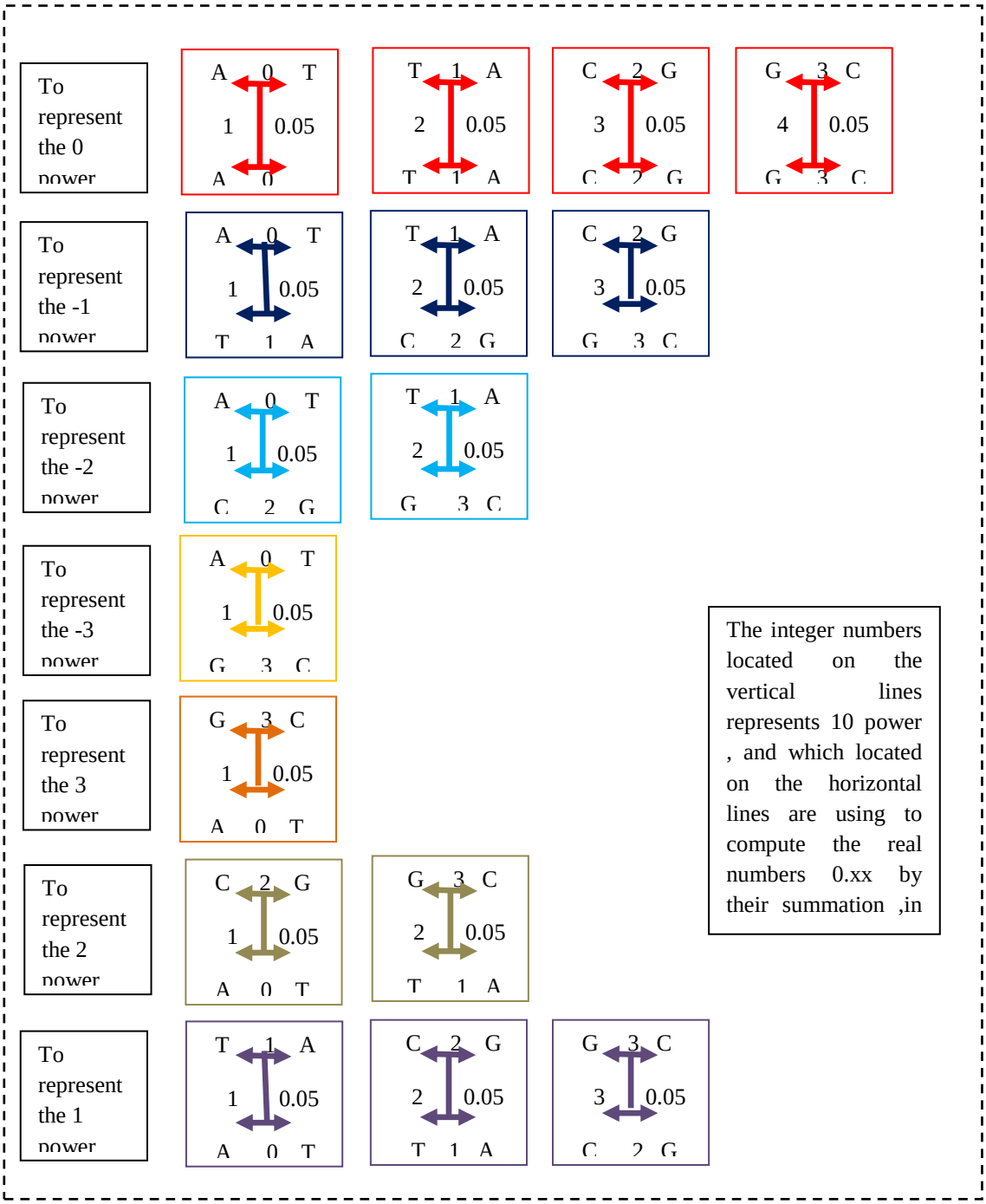

Figure 2 : The mathematical model for the genetic bonds

4 An example

Secret information(message) :"SECURE CHANNEL COMMUNICATION"
mRNA sequence[1] = AUG AAC GGC UCG CCC GGU CUG GUC UAC AUG GAG UCG GUG GCC AAC CUG CUG GAG GAG CCC UAA GGU AUC……

Stage 1:

Letters	C	N	E	A	I	M	O	U	H	L	R	S	T
freq.	4	4	3	2	2	2	2	2	1	1	1	1	1
Codons	GGC	GUC	GUG	CUG	UAC	GAG	UAU	AUC	AAC	UCG	CCC	GGU	GCC
freq.	4	4	4	3	3	3	3	3	2	2	2	2	2
Real No.s	0.75	3.5	1.25	0.33	3	4.33	5	7	4	6	9	9.5	10

Stage 2:

Real No.s	$0.75*10^0$	$0.34*10^1$	$0.12*10^1$	$0.33*10^0$	$0.3*10^1$	$0.43*10^1$	$0.5*10^1$	$0.7*10^1$	$0.4*10^1$	$0.6*10^1$	$0.9*10^1$	$0.95*10^1$	$0.1*10^2$

Stage 3 :

Message letters ➔ S E C U R E C
its Real No.s ➔ $0.95*10^1$ $0.12*10^1$ $0.75*10^0$ $0.7*10^1$ $0.9*10^1$ $0.12*10^1$ $0.75*10^0$

message letters ➔ H A N N E L C
its real No.s ➔ $0.4*10^1$ $0.33*10^0$ $0.34*10^1$ $0.34*10^1$ $0.12*10^1$ $0.6*10^1$ $0.75*10^0$

message letters ➔ O M M U N I C
its real No.s ➔ $0.5*10^1$ $0.43*10^1$ $0.43*10^1$ $0.7*10^1$ $0.34*10^1$ $0.3*10^1$ $0.75*10^0$

message letters ➔ A T I O N
its real No.s ➔ $0.33*10^0$ $0.1*10^2$ $0.3*10^1$ $0.5*10^1$ $0.34*10^1$

Stage 4:

letter	Real No.	DNA part Head	DNA part tail	DNA sub-strand
S	$0.95*10^1$	GC	GC(0.15)+GC(0.4)+GC(0.4)	GCGCGC
E	$0.12*10^1$	TA	TA(0.1)	TA
C	$0.75*10^0$	GG	GG(0.5)+TT(0.2)+AA(0.05)	GGTTAA
U	$0.7*10^1$	GC	CT(0.25)+CT(0.25)+TA(0.1)+TA(0.1)	CTCTTATA
R	$0.9*10^1$	GC	GC(0.4)+GC(0.4)+TA(0.1)	GCGCTA
E	$0.12*10^1$	TA	TA(0.1)	TA
:	:	:	:	:

1 * These codons series chosen from primeval nucleus (E.coli) . EBI web site.

Note :When we are combining all sub strands in one double strand ,we will separate their by mutates .

The information amount measure (entropy) illustrates the deference between the sequence of the message letters and the sequence of genetic bases were contents by DNA strand. The deference shows that information amount in DNA strand larger than of its in the message(as shown in figure 3).

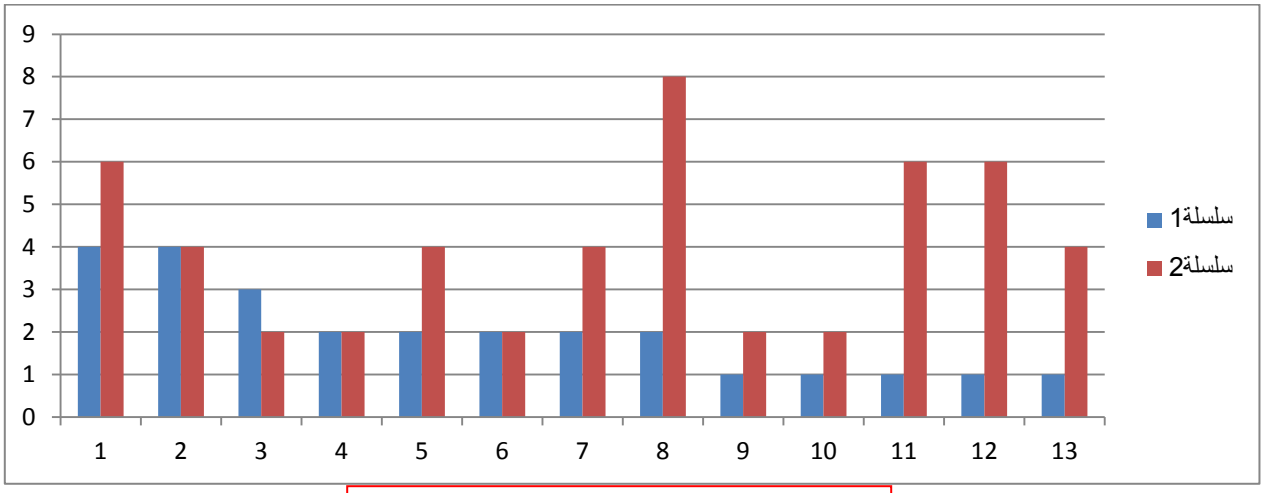

Figure 3:the entropy indicator of two sequences.
Series1 :frequency of message letters

5 Discussion

In cryptology or embedding works , the complexity of any suggested algorithm is an important feature for it, so, our technique to be distinguished by some properties make it satisfactory work, some of these are:

1 Essentially , it can be considered as a real simulation of what to do in biological world of central dogma processes.

2 There are many ways when DNA composed of mathematical bonds to get variety in the establishment.

3 The produced codes consider as a variable length codes.

4 The resulted digital DNA can be constructed in the laboratory of biology.

We can benefit from numerous steps, inside the stages, to increase the complexity of the computing power, for strong encryption purpose ,this step begins with numbers choosing and distributing in mathematical bonds model, the other is selecting secured codons, that are opposite to the message letters , eventually, the sub strands can be combined to get one DNA strand ,through manners are swelling the complexity potency .

The technique has come to fill. Beginning with the objectives of abnormal communication security , uncommon hiding environment ,and mathematical propositions used to achieve these objectives, the paper gives us a panoramic view of the fruits of experiences and efforts of public research in heredity genetics. The title says it all; from the mundane objective of having a secure communication the very first time you join someone to the

possibilities of digital money and cryptographically secure elections, this is where you'll find it.

6 References

Ban Ahmed Mitras and Adeeba Khaboo ,2013,"proposed steganography approach using DNA properties ",international journal of information technology and business management ",vol.4,no.1,pp:96-102,june.

Bruce Sheiener ,1997, "Applied cryptography ",second edition ,USA.

Feng B.S.J. and Potkonjak M.,2002 ,"data hiding in DNA",international workshop on information hiding,vol.4,pp:373-386,june.

Hayam Mouse ,Kamel Moustafa ,Waiel Abdel-Wahed, and Mahiy Hadhoud,2011,"data Hiding based on contrast mapping using DNA medium ",the international arab journal of information technology ,vol.8 no.2, pp:147-154 ,April .

Heitorsilverio and Leonardo Magalhaes Cruz ,2011,"computational biology and applied bioinformatics", InTech printing , Croatia .

Henk C.A. van Tilborg,2007, "Fundamentals of cryptology ,professional reference and interactive tutorial",kluwer academic publishers , London.

Motamemi H. and etal.,2007 "labeling method in steganography",world acad Sci. Eng. technology Issue 30,pp:349-354 .

Paolella P.,1998,"Introduction to molecular biology",Mcgraw – Hill companies Inc., new York.

Watson J.D. and Crick F.H.,1953 ,"molecular structure of nucleic acids:a structure for deoxyribose nucleic acid nature" ,171:737-738, USA.

Prased M.S. and etal.2011,"a Novel information hiding technique for security by using image steganography", journal of theoretical and applied information technology,vol.8,no.1,pp:35- 39,February .

Rohani Binti Abu Bakar and Junzo Watada,2008,"DNA computing and its applications:survey",waseda university ,ICIC express letters ,vol.2 ,no.1, march .

Samir Kuwar Bandyopodhyay , and Suman Chakraborty, 2011"Image hiding in DNA sequence using arithmetic encoding",journal of global research in computer science ,vol.2,no.4,pp:167- 171,April.