Using DNA to Encode Text Files

Yassin H. Ismail **Najla B. Ibrahim**
College of Computer Science and Mathematics
University of Mosul, Iraq

Accepted on: 18/9/2012 **Received on: 11/6/2012**

ABSTRACT

Modern studies focused on Deoxyribonucleic acid (DNA) because that DNA have several important features including the random nature of the sequence of nitrogenous bases consisting the acid and large storage capability of the information that led to it's usage in the field of encryption where the appearance of a new branch which is encryption of DNA . This research provided a new method to encrypt text files using DNA were building a set of coding tables and using them to obtain the cipher text in DNA form , also used a set of transposition cipher methods for the purpose of increasing the security of the resulted cipher text .

Keywords: DNA , Encode.

استخدام الـ DNA في تشفير الملفات النصية

ياسين حكمت إسماعيل ***نجلاء بديع ابراهيم***

كلية علوم الحاسوب والرياضيات، جامعة الموصل

تاريخ استلام البحث : 11\6\2012 **تاريخ قبول البحث : 18\9\2012**

الملخص

اهتمت الدراسات الحديثة بالحامض النووي الرايبي منقوص الأوكسجين (DNA) وذلك لامتلاكه العديد من الميزات المهمة منها الطبيعة العشوائية لتسلسل القواعد النتروجينية المكونة للحامض و قابلية الخزن الكبيرة للمعلومات والتي أدت إلى استخدامه في مجال التشفير إذ ظهر فرع جديد وهو تشفير الـ DNA . في هذا البحث قدمت طريقة جديدة لتشفير الملفات النصية باستخدام الـ DNA فقد تم بناء مجموعة من جداول الترميز واستخدامها للحصول على نص مشفر يكون بصيغة حامض الـ DNA كذلك فقد استخدمت مجموعة من طرائق التشفير الأبدالية لغرض زيادة الأمنية في النص المشفر الناتج .

الكلمات المفتاحية : DNA، التشفير .

1. المقدمة

إن تشفير الحامض النووي الرايبي منقوص الأوكسجين (DNA) حظي بأهتمام كبير نظرا" لقابلية الخزن الكبيرة للحامض إذ أن غراما" واحدا" منه له القابلية على خزن بيانات تقدر بـ 10^8 (Tera Bytes) هذه القابلية لخزن المعلومات تفوق كل وسائط الخزن المعروفة (الكهربائية ، المغناطيسية ، الضوئية) [7] [9] .

الأحماض النووية هي مركبات كيميائية معقدة التركيب توجد في جميع الأحياء وهي ذات أهمية كبيرة لها ، إن جزيئات الحامض النووي الرايبي منقوص الأوكسجين مؤلفة من عدد كبير من الوحدات الأصغر تعرف بالنيوكليوتيدات (Nucleotides) [1] .

يتألف كل جزيء من النيوكليوتيد من ثلاث جزيئات أبسط مرتبط بعضها ببعض مباشرة وهي [1] [3],[4] :

1. قاعدة نتروجينية : وهي مركب حلقي يحتوي على النتروجين بالإضافة إلى الكاربون والهيدروجين والأوكسجين (عدا الأدنين إذ لا يحتوي على الأوكسجين) و يوجد منها نوعان هما :

أولا : البريميدينات : وتتكون من حلقة واحدة وتشمل القواعد الآتية :

أ.الثايمين T ، ب.السايتوسين C

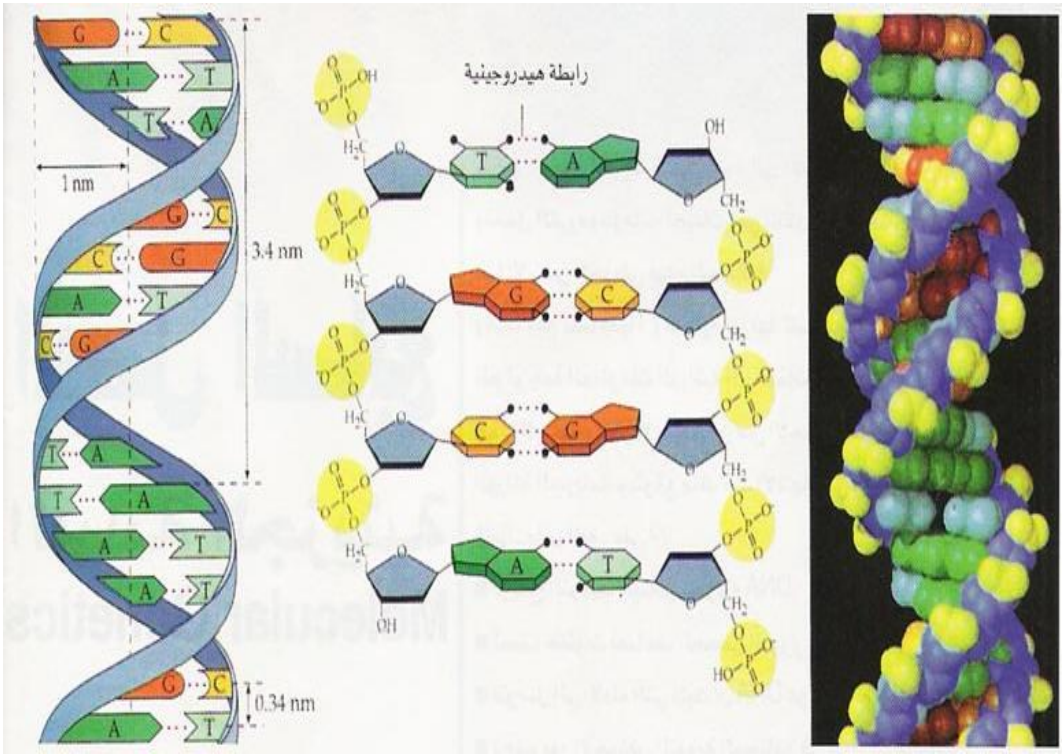
ثانيا : البيورينات : وتتكون من حلقتين وتشمل القواعد الآتية :

أ.الأدنين A ، ب.الكوانين G

2. سكر خماسي الكاربون : وهو سكر الرايبوز منقوص الأوكسجين الذي يختلف عن الرايبوز بفقدانه ذرة أوكسجين واحدة ، وصيغته الجزيئية هي ($C_5H_{10}O_4$) .

3. حامض الفوسفوريك .

قدم واطسن وكريك [1] [4] عام 1953 أنموذجا" للحامض النووي DNA مؤلفا" من سلسلتين أو شريطين ملتفين على هيئة سلم حلزوني ترتبط فيه إحدى القواعد النتروجينية في أحد شقي الحلزون مع القاعدة النتروجينية للشق الآخر بواسطة أواصر هيدروجينية . إن إرتباط القواعد النتروجينيبة بين الشقين لا يكون عشوائيا" بل مقيدا" ، فالأدنين في أحد الشريطين يرتبط دائما مع الثايمين في الشريط الآخر باصرتين هيدروجينيتين ويرتبط السايتوسين في أحد الشريطين مع الكوانين في الشريط الاخر بثلاث أواصر هيدروجينية وكما موضح بالشكل رقم (1) .



الشكل رقم (1) يوضح نموذج العالمان واطسن و كريك

2. الدراسات السابقة

أول من أستخدم الحامض النووي الرايبي منقوص الأوكسجين (DNA) في مجال الحساب هو العالم أديلمان [10] عام 1994 لحل مشكلة إيجاد أفضل مسار ، إذ وجد بأن الحامض يحتوي على خاصية المعالجة المتوازية والتي توفر سرعة عالية جدا إذا ما تم استغلالها في المجالات الحسابية . في عام 1995 قام بونييه و آخرون [6] باستخدام حامض الـ DNA لكسر نظام تشفير البيانات القياسي (DES) . بعد ذلك ظهرت العديد من المحاولات [9] [12] [11] [5] [8] في بناء أنظمة تشفير للبيانات تعتمد على حامض الـ DNA إذ استخدمت خاصية عشوائية تسلسل القواعد النتروجينية للحامض وعدها مفتاحا" لنظام تشفير المرة الواحدة (One-time Pad

System) . نظريا" نظام التشفير الذي يعتمد خاصيتي عشوائية مفتاح التشفير واستخدامه لمرة واحدة لا يمكن أن يكسر [2] .

إن استخدام الحامض النووي الرايبي منقوص الأوكسجين في مجال التشفير يوفر مستويين من الحماية للبيانات المشفرة ، الأول يكمن في صعوبة إجراء واستخدام العمليات والتقنيات البايولوجية والمستوى الثاني يتعلق بصعوبة حل و تحليل العمليات الرياضية المستخدمة في تشفير البيانات [5] .

3. أنواع أنظمة التشفير

يمكن تقسيم أنظمة التشفير من حيث آلية تعاملها مع أحرف النص الواضح لغرض الحصول على النص المشفر إلى نوعين [2] :

1- أنظمة التشفير الأبدالية (Transposition Cipher System) : يتم في هذا التشفير أعادة ترتيب حروف الرسالة الواضحة بحيث تبقى بدون تغيير فقط يتم تغيير مواقعها ضمن النص ، إذ يتم تشفير " SEND HELP SOON" إلى "SLEPNSDOHOEN" . أي أن إحصائية أحرف النص الواضح تكون مساوية لإحصائية النص المشفر فمثلا يحتوي النص الواضح في المثال أعلاه على حرفين "N" وكذلك النص المشفر الناتج .

2- أنظمة التشفير التعويضية (Substitution Cipher System) : هنا يتم استبدال حروف النص الواضح بحروف أخرى أو أعداد أو رموز ، فمثلا" يتم تشفير كلمة "COMPUTER" إلى "XRSYMHZK" .

4. هدف البحث

يهدف البحث إلى تقديم طريقة جديدة لتشفير الملفات النصية بالاعتماد على استخدام فكرة الحامض النووي الرايبي منقوص الأوكسجين(DNA) في عملية التشفير . تم بناء مجموعة من جداول الترميز (طرق التشفير التعويضية) واستخدامها مع طرائق التشفير الأبدالية لغرض الحصول على النص المشفر والذي يكون بصيغة حامض الـ DNA . الطريقة المقترحة تدمج بين مفهوم أنظمة التشفير الابدالية والتعويضية و استخدام الجداول بوصفها مفتاحا" سريا" بين الطرفين و بالتالي الحصول على درجة عالية من الأمنية للنص المشفر الناتج والذي يكون عبارة عن تسلسل من القواعد النتروجينة العشوائية المكونة لحامض الـ DNA .

5. الطريقة المقترحة

يطلق على كل ثلاث قواعد نتروجينية بالكودون (Codon) وبما أنه هنالك أربعة نيوكليوتيدات تدخل في تركيب حامض الـ DNA وهي (T,A,C,G) إذن هنالك أربعة وستون كودونا" مختلفا" يمكن الحصول عليه من هذه القواعد وكما موضح بالجدول رقم (1) [3] [4] .

يتألف النص الواضح المكتوب باللغة الانكليزية من تتابعات مختلفة لأحرف اللغة الانكليزية والتي عددها ستة وعشرون حرفا" (A-Z) . يتم ترميز أحرف النص الواضح إلى كودونات بحيث يرمز كل حرف إلى كودون معين اعتمادا" على موقع الحرف ضمن النص الواضح حيث أن الحرف الذي يقع في موقع فردي (يشار له بالحرف الصغير.Small Char) يرمز إلى كودون مختلف عن ترميز نفس الحرف ولكن يقع في موقع زوجي (يشار له بالحرف الكبير.Capital Char) ضمن سلسلة أحرف النص الواضح وكما هو موضح بالجدول رقم (2).

يتم الاتفاق بين الجهة المرسلة والمستلمة على سلسلة محددة من الحامض النووي الرايبي منقوص الأوكسجين (DNA) وهذه المعلومات متوفرة ضمن قواعد بيانات في مراكز عالمية مختصة في الهندسة الوراثية و

دراسة عمل الجينات و كذلك في العديد من مواقع الانترنت مثل (EMBL [13] , NCBI [14] , GENBANK [15]) ، حيث تمثل هذه السلسلة المتفق عليها جزءا من المفتاح السري ، تجري عملية ترميز للقواعد النتروجينية لسلسلتي الحامض (الأولى تمثل النص الواضح بعد ترميزه إلى كودونات والثانية هي سلسلة الحامض المتفق عليها) ويكون ناتج ترميز القاعدتين عبارة عن حرف من الأحرف الانكليزية وحسب الجدول رقم (3) وكما نلاحظ بالجدول هنالك 16 احتمالا" لترميز القاعدتين .

بعد الحصول على سلسلة الأحرف تحول تلك الأحرف إلى النظام الثنائي وبالاعتماد على تسلسلها ضمن أبجدية اللغة الانكليزية وبما أنه هنالك 26 حرفا" انكليزيا" إذن كل حرف يرمز بخمس وحدات (Digits) في النظام الثنائي ، بعد ذلك يتم استخدام إحدى طرائق التشفير الابدالية لإعادة ترتيب تسلسل الأرقام الثنائية .

ترمز السلسلة الثنائية الناتجة بعد عملية الإبدال إلى القواعد النتروجينية بالاعتماد على الجدول رقم (4) ، تستخدم طريقة تشفير أبدالية أخرى لإعادة ترتيب تسلسل القواعد النتروجينية الناتجة وبالتالي الحصول على سلسلة عشوائية من القواعد النتروجينية والتي تمثل النص المشفر .

يمكن توضيح خطوات عملية التشفير و فك التشفير بالمخطط الانسيابي في الشكلين (2) و (3) .

جدول رقم (1) يوضح عملية ترميز الكودونات الثلاثية من القواعد النتروجينية الأربعة

القاعدة الثانية

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

جدول رقم (2) ترميز أحرف النص الواضح إلى كودونات حسب موقعها

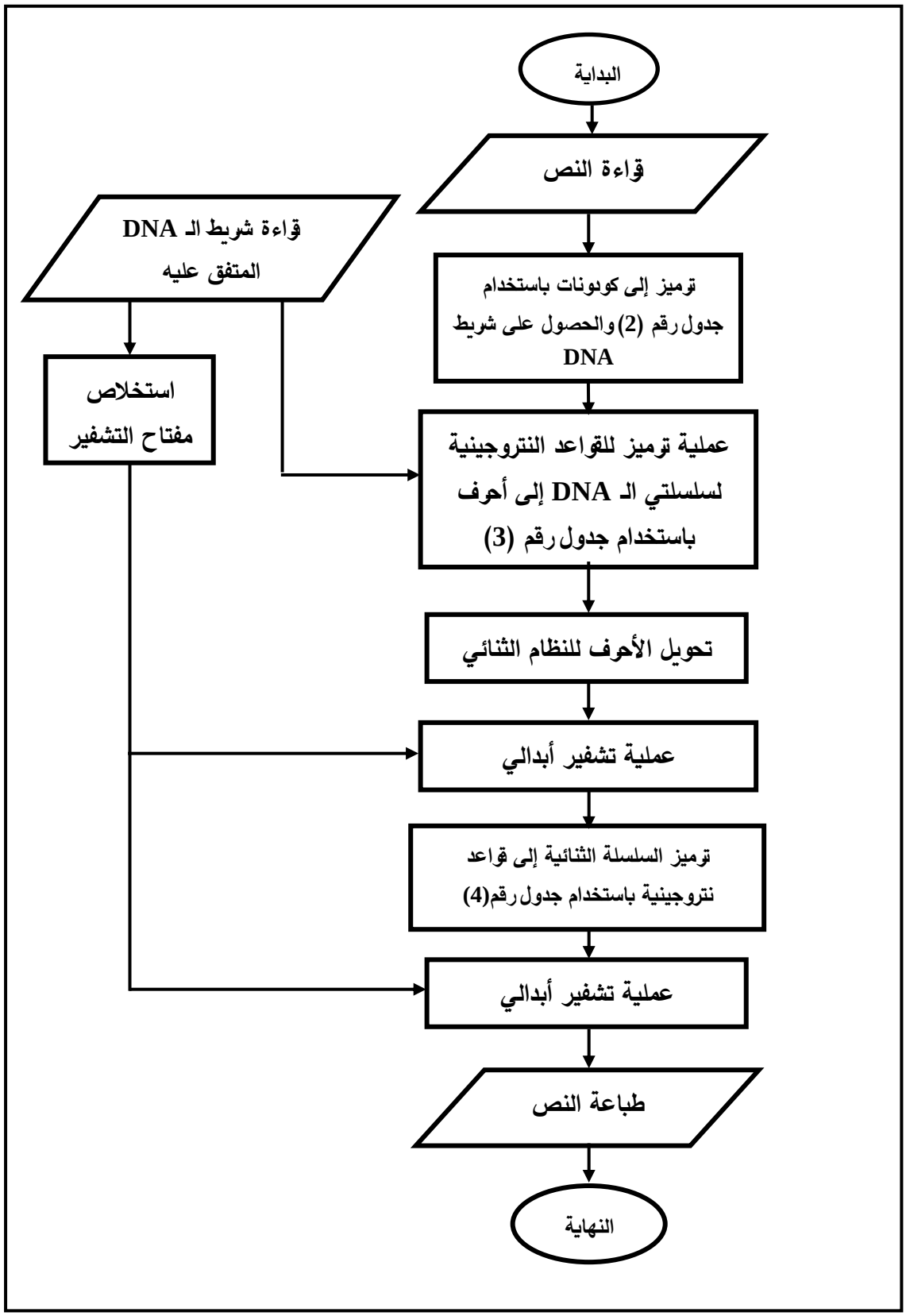
الحرف الفردي	الكودون	الحرف الفردي	الكودون	الحرف الزوجي	الكودون	الحرف الزوجي	الكودون
a	TTC	n	TTA	A	TTG	N	TCT
b	TCC	o	TCA	B	TCG	O	TAT
c	TAC	p	TGT	C	TGC	P	TGG
d	CTT	q	CTC	D	CTA	Q	CTG
e	CCT	r	CCA	E	CCG	R	CAT
f	CAC	s	CAA	F	CAG	S	CGT
g	CGC	t	CGA	G	CGG	T	ATT
h	ATC	u	ATA	H	ATG	U	ACT
i	ACC	v	ACA	I	ACG	V	AAT
j	AAC	w	AAG	J	AGT	W	AGC
k	AGA	x	AGG	K	GTT	X	GTC
l	GTA	y	GTG	L	GCT	Y	GCC
m	GCA	z	GCG	M	GAT	Z	GAC

جدول رقم (3) عملية ترميز قاعدتين نتروجينية إلى حرف انكليزي

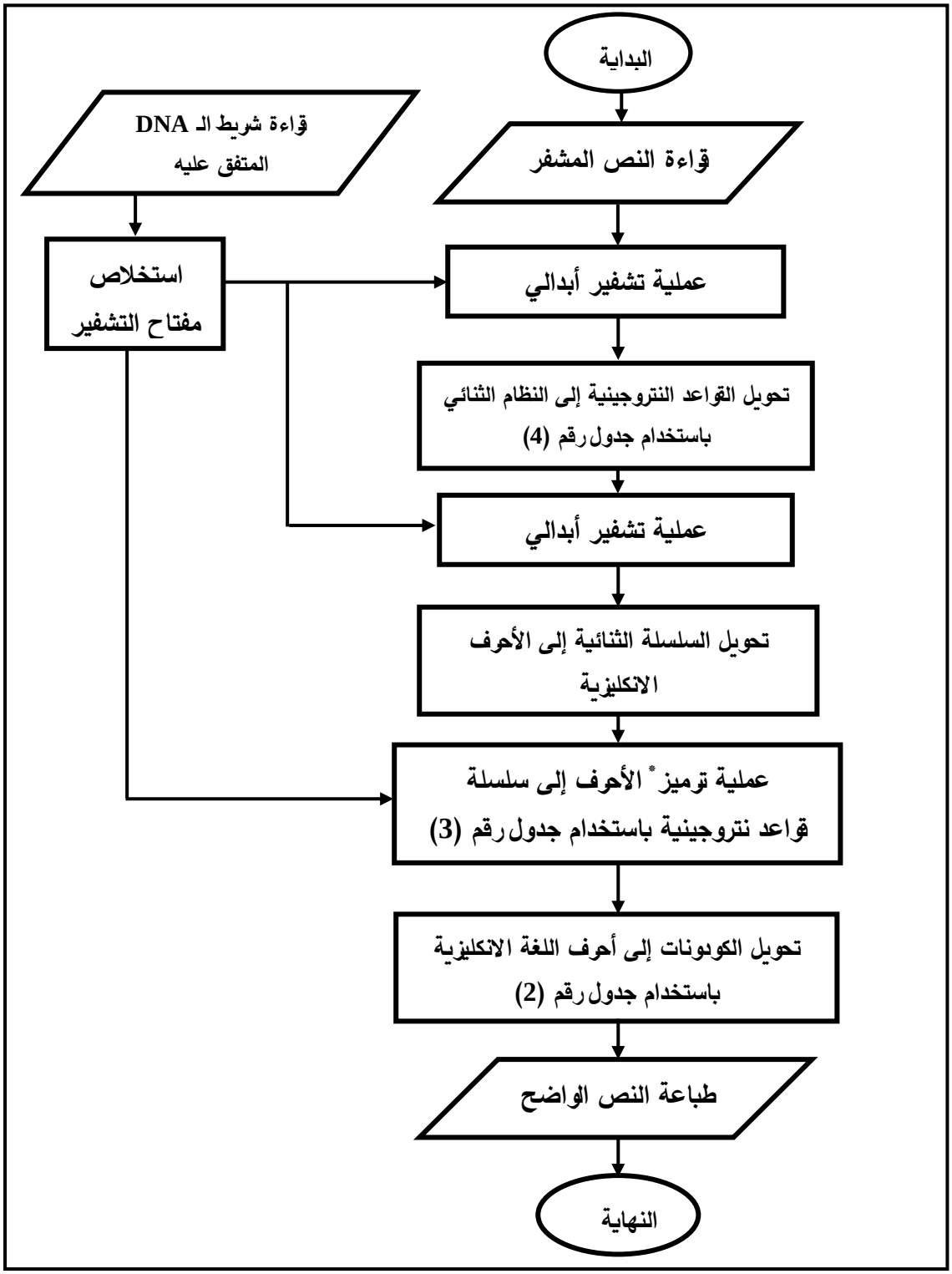
سلسلة DNA ناتجة من ترميز النص الواضح	سلسلة DNA متفق عليها بين الطرفين	الحرف الانكليزي المرمز للقاعدتين	سلسلة DNA ناتجة من ترميز النص الواضح	سلسلة DNA متفق عليها بين الطرفين	الحرف الانكليزي المرمز للقاعدتين
T	T	B	A	T	D
T	G	H	A	G	I
T	A	K	A	A	R
T	C	M	A	C	P
G	T	S	C	T	X
G	G	V	C	G	W
G	A	Y	C	A	Z
G	C	E	C	C	F

جدول رقم (4) ترميز الأرقام الثنائية إلى قواعد نتروجينية

الأرقام الثنائية	القاعدة المقابلة لها
00	A
01	C
10	G
11	T



الشكل رقم (2) المخطط الانسيابي لعملية التشفير للطريقة المقترحة



الشكل رقم (3) المخطط الانسيابي لعملية فك التشفير للطريقة المقترحة

* حيث لدينا الأحرف و سلسلة القواعد النتروجينية لشريط الـ DNA المتفق عليه فنحصل على الطرف الثالث بسهولة و حسب الجدول رقم (3) .

6. مثال توضيحي للطريقة

لتوضيح آلية عمل الطريقة المقترحة سوف يتم تشفير نص معين بالاعتماد على المخطط الانسيابي في الشكل 2 ، علما" بأن عملية فك التشفير تتم بالخطوات نفسها ولكن بصورة معكوسة وبالأعتماد على المخطط الأنسيابي في الشكل 3 .

عملية التشفير : تتم عملية التشفير من خلال الخطوات التالية :

1. قراءة النص الواضح ونفترض أنه "send help soon" .

2. قراءة شريط الـ DNA المتفق عليه بين المرسل والمستلم ، نفترض أن الاتفاق لبكتريا (أيكولاي E.coli) مأخوذة من موقع EBI حيث أن الشريط المقروء هو :

G1=ATGTGCGAAAAAAACGCCTATTTGCTCCTGTGGGAAGGACCATGAACTA
AGGCGCCT…

3. ترميز أحرف النص الواضح إلى كودونات وحسب الجدول رقم 2 وكما موضح :

الأحرف	s	e	n	d	h	e	l	p	s	o	o	n
موقعها*	o	e	o	e	o	e	o	e	o	e	o	e
الكودون المرمز	CAA	CCG	TTA	CTA	ATC	CCG	GTA	TGG	CAA	TAT	TCA	TCT

نلاحظ من عملية الترميز أن الحرف n الواقع في الموقع الفردي يختلف ترميزه عن الحرف n الواقع بالموقع الزوجي وكذلك الحرف o في حين أن الحرفين s و e يقعان في نفس المواقع لذلك يكون لهما نفس الكودون المرمز ، مما سبق نجد أن اعتماد موقع الحرف في اختيار الكودون المرمز وفرت مستوى عاليا" من العشوائية و قضت على مشكلة إحصائية الأحرف الانكليزية حيث في مواقع معينة للأحرف ضمن النص الواضح يكون نفس الكودون المرمز وفي مواقع أخرى يكون الترميز إلى كودون مختلف .

* نرمز للأحرف التي تقع بالمواقع الفردية ضمن سلسلة النص الواضح بالحرف o بينما يرمز للأحرف الواقعة بالمواقع الزوجية بالرمز e .

4. ترميز القواعد النتروجينية لسلسلتي الحامض النووي إلى أحرف انكليزية باستخدام جدول رقم 3 وكما موضح :

سلسلة الأحرف الانكليزية الناتجة	سلسلة الـ DNA المتفق عليها	سلسلة الـDNA الناتجة من الخطوة 3	سلسلة الأحرف الانكليزية الناتجة	سلسلة الـ DNA المتفق عليها	سلسلة الـDNA الناتجة من الخطوة 3
S	T	G	Z	A	C
K	A	T	D	T	A
D	T	A	I	G	A
B	T	T	X	T	C
S	T	G	W	G	C
V	G	G	E	C	G

F	C	C	H	G	T
D	T	A	K	A	T
P	C	A	R	A	A
M	C	T	Z	A	C
D	T	A	K	A	T
H	G	T	R	A	A
B	T	T	R	A	A
W	G	C	K	A	T
I	G	A	F	C	C
H	G	T	W	G	C
Z	A	C	F	C	C
K	A	T	E	C	G

5. تحويل الأحرف إلى سلسلة أرقام ثنائية حيث أن كل حرف يرمز إلى خمس وحدات (Digits) في النظام الثنائي وبالاعتماد على تسلسلها ضمن هجائية الأحرف الإنكليزية (A-Z) وكما هو موضح في الجدول التالي :

الأرقام الثنائية	تسلسلها في الأبجدية	الأحرف	الأرقام الثنائية	تسلسلها في الأبجدية	الأحرف
10011	19	S	11010	26	Z
01011	11	K	00100	4	D
00100	4	D	01001	9	I
00010	2	B	11000	24	X
10011	19	S	10111	23	W
10110	22	V	00101	5	E
00110	6	F	01000	8	H
00100	4	D	01011	11	K
10000	16	P	10010	18	R
01101	13	M	11010	26	Z
00100	4	D	01011	11	K
01000	8	H	10010	18	R
00010	2	B	10010	18	R
10111	23	W	01011	11	K
01001	9	I	00110	6	F
01000	8	H	10111	23	W
11010	26	Z	00110	6	F
01011	11	K	00101	5	E

و بالتالي تكون سلسلة الأرقام الثنائية هي :

1101000100010011100010111001010100001011100101001011010010111
0010100100101100110101110011000101100110101100100000101001110110001100
0100100000110100100010000000101011101001010001101001011

6. أجراء عملية تشفير إبدالي بإحدى الطرائق بأعتماد مفتاح تشفير مستنتج من سلسلة الـ DNA المتفق عليها وذلك لغرض زيادة العشوائية في السلسلة الثنائية فتكون السلسلة الناتجة هي :

1100110001011001101000010111001010010100100000010100111011000110001000
1100101010101001011110111001000001100000110100101000101001001011001101
1101001010011010001000100111000101011010010 11

7. تحويل السلسلة الثنائية إلى سلسلة حامض نووي أي سلسلة من القواعد النتروجينية بأستخدام جدول رقم 4 فتكون سلسلة القواعد الناتجة هي :

TATACCGCTCACCAGAGTATATCACATCACACATACCTCAT

8. أجراء عملية تشفير أبدالية لسلسلة القواعد النتروجينية و بأستخدام نفس مفتاح التشفير المستخدم في الخطوة 6 فتكون السلسلة الناتجة هي :

ATATCACCGCTCATACCTTATACACCACACATAGAATC GTC

من خلال ملاحظة سلسلة القواعد النتروجينية الناتجة والتي تمثل لنا النص المشفر الناتج يتبين درجة العشوائية العالية والكفاءة للطريقة المقترحة ، علما" بأن سلسلة القواعد النتروجينية الناتجة من الممكن أن تصنع من قبل شركات خاصة تقوم بعملية تكوين لحامض نووي رايبي منقوص الأوكسجين DNA وحسب التسلسل المرغوب فيه وبالتالي وخاصة مع الرسائل الطويلة والعالية الأمنية0 ممكن استخدام شريط الـ DNA وسطا" لنقل بيانات سرية .

7. الاستنتاجات

في هذا البحث تم تقديم طريقة جديدة لتشفير الملفات النصية إذ تم أستخدام الحامض النووي الرايبي منقوص الأوكسجين (DNA) في ذلك . تم بناء مجموعة من جداول التعويض و استخدامها مع الطرائق الأبدالية في عملية التشفير . وفرت طريقة التشفير المقترحة درجة عالية من الأمنية والعشوائية للنص المشفر الناتج . يحتوي الحامض النووي الرايبي منقوص الأوكسجين (DNA) العديد من الميزات المهمة منها قابلية الخزن العالية و سرعة تنفيذ العمليات الحسابية بالإضافة إلى الكلفة القليلة في عمليات الحساب والخزن هذه الميزات المهمة يمكن استغلالها في العديد من المجالات والاستعاضة عن رقاقة السليكون بشريط الـ DNA .

المصادر

[1] الجلبي قصي عبد القادر، (1991) ، "الأحماض النووية" ، دار الحكمة للطباعة والنشر ، الموصل – العراق.

[2] الحمامي علاء حسين ، (2007)، "تكنولوجيا أمنية المعلومات وأنظمة حمايتها" ، الطبعة الأولى ، دار وائل للنشر ، عمان – الاردن.

[3] تاج الدين سعد جابر، (2000) "علم الوراثة "،الطبعة الثانية، جامعة البصرة، البصرة– العراق.

[4] شكارة مكرم ضياء ، (2000)"علم الوراثة "،الطبعة الاولى، دار المسيرة ، عمان – الاردن.

[5] Beenish Anam , and et al , (2010), " Review on the Advancements of DNA Cryptography " , SKIMA Paro, Bhutan.

[6] Dan Boneh, Christopher Dunworth, and Richard J. Lipton ,(1996), " Breaking DES using a molecular Computer " , Discrete Mathematics and Theoretical Computer Science , American Mathematical Society.

[7] G. Cui, Y. Liu, and X. Zhang, (2006) , "New direction of data storage: DNA molecular storage technology," Computer Engineering and Application, vol. 42, no. 26, pp. 29–32.

[8] Guangzhao Cui , and et al. , (2008), " An Encryption Scheme Using DNA Technology " , IEEE 37 BIC-TA .

[9] J. Chen , (2003), "A DNA-based, biomolecular cryptography design", IEEE International Symposium on Circuits and Systems (ISCAS) .

[10] L. M. Adleman. , (1994), "Molecular computation of solutions to combinatorial problems". Science.

[11] Monica Brdoa , Olga Tornea, (2010) , " DNA Secret Writing Techniques " , IEEE.

[12] Xing Wang, Qiang Zhang ,(2009), " DNA computing-based cryptography " , IEEE .

[13] http://www. EMBL.com.

[14] http://www. NCBI.com.

[15] http://www. GENBANK.com.