Journal of Engineering and Technology Research Vol. 3(5), pp. 148-154, May 2011
Available online at http:// www.academicjournals.org/JETR
ISSN 2006-9790 ©2011 Academic Journals

Full Length Research Paper

Deoxyribonucleic acid (DNA) as a hypothetical information hiding medium: DNA mimics basic information security protocol

Okunoye .O. Babatunde

Department of Pure and Applied Biology, Ladoke Akintola University of Technology, P. M. B. 4000, Ogbomoso, Osun State, Nigeria. E-mail: babatundeokunoye@yahoo.co.uk. Tel: +234-08036296145

Accepted 12 April, 2011

While the computational capabilities of deoxyribonucleic acid (DNA) are now fairly known, with the building of DNA computers which utilize DNA and other sub-cellular molecules as input and output, little seems to have been investigated on the information security properties of DNA in computer science and information technology terms. The author investigated DNA as a hypothetical information hiding molecule, which employs the principles of steganography to obscure information that might be of importance to a hypothetical attacker. By employing the combinations and permutations of bases in DNA nucleotide sequences as substitutes for an English letter, the author shows that in addition to the computational capabilities of DNA, it also mimics basic information hiding and information security protocol.

Key words: Deoxyribonucleic acid (DNA) computing, information hiding, steganography, information security, biotechnology, bioinformatics.

INTRODUCTION

The recently emerging field of deoxyribonucleic acid (DNA) computing has witnessed the execution of calculations and the design of devices using sub-cellular molecules including DNA as input and output (Adleman, 1994; Quyang et al., 1997; Pirrung et al., 2000; Sakamoto et al., 2000; Wang et al., 2001; Benenson et al., 2001; Stojanovic et al., 2002; Benenson et al., 2003; Stojanovic and Stefanovic, 2003 a and b; Okamoto et al., 2004; Margolin and Stojanovic, 2005). Deoxyribonucleic acid alongside other sub-cellular molecules, have been used to perform routine calculations, while the cell is now seen as a computation (Regev and Shapiro, 2002). DNA and molecular computations have shown, in computer science terms, the computational capacity of DNA, even though detailed study of notable microbial DNA have revealed purely computational methods (Miller et al., 2003). An aspect of DNA computing the numerous literature in the expansive field, as cited previously, have laid less emphasis on the information hiding possibilities of the DNA molecule in computer science and information security terms, as distinct from the natural coding mechanisms in the field of molecular biology. In molecular biology, deoxyribonucleic acid is known to naturally encode for twenty amino acids in sixty-four triplet bases or codons (Weaver, 2005).

These form the basis of the natural transmission of information from generation to generation in humans and all living things. There are ten nucleotide bases per turn of a DNA helix, which comprises of a sequence of ten bases. All these and other delicate arrangements in DNA structure points to subtle information processing which makes transcription, translation and protein formation possible. The objective of this paper is to examine the structure of DNA as a hypothetical information hiding medium, in computer science terms. Information hiding is a young and rapidly evolving field with many techniques which include steganography, copyright marking, fingerprinting and water marking (Petitcolas et al., 1999). Many information hiding techniques however can trace their roots back to anquity (Petitcolas et al., 1999) and have extensive applications among the military, intelligence agencies, and law enforcement.

Here, we examine hypothetical DNA steganography, as a sub-discipline of information hiding. Whereas classical cryptography is concerned about hiding the content of messages, steganography deals with concealing their existence (Anderson and Petitcolas, 1998).

Considering part of a work of cryptography (Sutton and Rubin, 2009), complicit in one of the historical events in international relations, intercepted and decoded by British Intelligence officers (The Zimmermann Telegram sent by the German Government to the Mexican Government during World War 1):

German Legation Jan 19, 1917 via Galveston
Mexico City

130 1342 13401 8501 115 3528 416 17214 6491 11310
18147 18222 21560 10247 11518 23667 13605 3494 14936
98092 5905 11311 10392 10371 0302 21290 5161 39695
23571 17504 11269 18276 18101 0317 0228 17694 4473

BERNSTORFF
Charge German Embassy
TELEGRAM RECEIVED
FROM 2nd from London # 5474

"We intend to begin on the first of February unrestricted marine warfare. We shall endeavour in spite of this to keep the United States neutral. In the event of this not succeeding, we make Mexico a proposal of alliance on the following basis…"

Many classical steganographic techniques date back to antiquity with Æneas the Tactician (Tacitus, 1990) inventing a number of such techniques. A detailed study of steganographic techniques, ancient and modern is discussed in Petitcolas et al. (1999). The general model of hiding data (Petitcolas et al., 1999) can be described as follows: The embedded data is the message one wishes to send secretly. It is usually hidden, in this case, in an innocuous message referred to as a cover-text, producing the stegotext. It is important to state that we do not imply that DNA actually hides any data of interest to a human observer. The paper aims to show that DNA can be viewed as a hypothetical stegotext, where the distribution of the combinations of its nucleotide bases per turns of DNA helix mimic the distribution of letters in a natural language such as the English language.

MATERIALS AND METHODS

The DNA used was Bacteriophage T4 DNA, obtained from GenBank, the institutional DNA depository, with accession number AF158101. With a complete genome sequence of 168,903 base pairs, Bacteriophage T4 represents the most understood model for modern genomics and proteomics (Miller et al., 2003), and has its study had revealed many insights and paradigms in molecular biology.

Advantage was taken of the mathematical combination of four nucleotides (adenine, guanine, thymine and cytosine) per turn of the DNA helix, which consist of ten bases (Nelson and Cox, 2000). Mathematically, there are 23 possible combinations and 296 permutations of four numbers with the sum of ten (Table 1).

The combinations of nucleotides in 3,183 helical segments of Bacteriophage T4 DNA (consisting of ten bases) were recorded (Table 2).

This represents 31,830 bases, complement 168,900' to 137,070' in the 5' to 3' direction. For the related complement, 21 nucleotide combinations were discovered; the combinations 0, 0, 0, 10 and 0, 0, 1, 9 did not occur. The frequencies of the nucleotide combinations were recorded while, similarly, the frequencies of 3,183 letters of the English language from the 5th chapter of a classic English literature text, Wuthering Heights (Figures 1 and 2) (Bronte, 1965) were recorded (Table 3). Considerable reductions in text are possible in the English language without losing information due to the statistical nature of the language and high frequency of certain words. This property is called redundancy (Shannon, 1949). Redundancy is of central importance in information security, and given that there was a disparity between the number of nucleotide combinations and English letters, we omit the letters C, Q, V, X, Z to bring the symbols to numerical parity of 21 pieces. Due to the fact that we assumed, hypothetically, that the nucleotide combinations are hiding information in the English language, we set up a simple substitution table between the nucleotide symbols and English letters, in order of increasing frequency (Table 4).

RESULTS

Consider the following sequence of helical combinations from T4 Phage DNA:

3': 0055 1234 0370 1144 4411 2611 2035 1414 1333
1342 1018 1117 2224 2242 0334 6004 1045 4051
2008 3016 0316 4132 1243 3007 8110 3115 5104
2323 2332 4060 2440 1216 1711 3340 4321 4006: 5'

This is comparable with a standard work of cryptography such as the Zimmermann Telegram (Sutton and Rubin, 2009). Interception of this traffic in a security situation could raise suspicion, especially if the interceptor is oblivious of their source. It might be argued however that the traffic is trivial for a capable attacker, as it is obvious the numbers are merely a mathematical operation (combinations) on the number set ($x: 0 \leq x \leq 10$); nevertheless the robustness of the traffic is its concealment within the cell protoplasm, encoded in the arrangement of nucleotide bases in DNA helices, making it a very uncanny example of "covered writing", hence steganography. Therefore, any attempt to retrieve the message by a hypothetical attacker will involve molecular biology procedures for DNA extraction, purification and analysis.

With the substitution of each respective nucleotide combination as depicted previously in the T4 phage DNA segment with the corresponding English letter, we obtain an embedded data containing 3,183 letters in the English language. Since we assume the stegotext, T4 phage DNA is concealing a message in English language, the probable word method (Shannon, 1949), is used to recover any embedded message within the stegotext. The probable words may be words or phrases expected

Table 1. Possible helical nucleotide combinations and permutations.

Nucleotide combination	Number of permutation	Permutation
0,0,0,10	4	00010,0010,01000,10000
0,0,1,9	12	0019, 0091,0109,0190,0901,0910,1009,1090,1900,1009,9001,9100
0,0,2,8	12	0028,0082,0208,0280,0802,0820,2008,2080,2800,8200,2080,8020
0,0,3,7	12	0037,0073,0307,0370,0703,0730,3007,3070,3700,7300,7003,7030
0,0,4,6	12	0046,0064,0406,0460,0604,0640,4006,4060,4600,6004,6040,6400
0,0,5,5	6	0055,0505,0550,5005,5050,5500
0,1,2,7	24	0127,0172,0217,0271,0712,0721,1207,1270,1702,1720,2017,2071 2107,2170,2701,2710,7012,7021,7102,7120,7201,7210,1027,1072
0,1,3,6	24	0136,0163,0316,0361,0613,0631,1036,1063,1306,1360,1603,1630, 3016,3061,3106,3160,3601,3610,6013,6031,6103,6130,6301,6310
0,1,1,8	12	0118,0181,0811,1018,1081,1108,1180,1801,1810,8011,8101,8110
0,2,2,6	12	0226,0262,0622,2026,2062,2206,2260,2602,2620,6022,6202,6220
0,2,3,5	24	0235,0253,0325,0352,0523,0532,2035,2053,2305,2350,2503,2530, 3025,3052,3205,3250,3502,3520,5023,5032,5203,5230,5302,5320
0,2,4,4	12	0244,0424,0442,2044,2404,2440,4024,4042,4204,4240,4402,4420
0,3,3,4	12	0334,0343,0433,3034,3043,3304,3340,3403,3430,4033,4303,4330
1,1,1,7	4	1117,1171,1711,7111
1,1,2,6	12	1126,1162,1216,1261,1612,1621,2116,2161,2611,6112,6121,6211
1,1,3,5	12	1135,1153,1315,1351,1513,1531,3115,3151,3511,5113,5131,5311
1,1,4,4	6	1144,1414,1441,4114,4141,4411
1,2,2,5	12	1225,1252,1522,2125,2152,2215,2251,2512,2521,5122,5212,5221
1,3,3,3	4	1333,3133,3313,3331
1,2,3,4	12	1234,1243,1324,1342,1423,1432,2134,2143,2314,2341,2413,2431, 3124,3142,3214,3241,3412,3421,4123,4132,4213,4231,4312,4321
2,2,3,3	6	2233,2323,2332,3223,3232,3322
2,2,2,4	4	2224,2242,2422,4222
0,1,4,5	24	0145,0154,0415,0451,0514,0541,1045,1054,1405,1450,1504,1540, 4015,4051,4105,4150,5014,5041,5104,5140,4501,4510,5401,5410
Total	296	

Table 2. Nucleotide combinations, permutations, frequencies and probabilities in 3,183 turns of T4 phage genome.

S/N	Combination		Permutation	Frequency	Probability
1	0,0,5,5	(X18)	6	5	0.0016
2	0,0,2,8	(X19)	1 2	7	0.0022
3	0,1,1,8	(X21)	1 2	11	0.0035
4	1,1,1,7	(X14)	4	15	0.0047
5	0,0,3,7	(X1 7)	1 2	16	0.0050
6	0,0,4,6	(X2 0)	12	23	0.0072
7	0,2,2,6	(X1 6)	1 2	63	0.0198
8	0,1,2,7	(X3)	24	76	0.0239
9	0,1,3,6	(X10)	24	109	0.0343
10	1,1,4,4	(X13)	6	119	0.0374
11	0,2,4,4	(X6)	1 2	1 35	0.0424
12	0,1,4,5	(X1)	24	140	0.0440

Table 2. Cont'd.

13	1,1,2,6	(X 8)	1 2	144	0.0453
14	0,3,3,4	(X12)	1 2	145	0.0456
15	1,3,3,3	(X9)	4	181	0.0569
16	2,2,2,4	(X15)	4	188	0.0591
17	1,1,3,5	(X7)	1 2	195	0.0613
18	0,2,3,5	(X11)	24	225	0.0707
19	1,2,2,5	(X4)	1 2	290	0.0911
20	2,2,3,3	(X2)	6	295	0.0927
21	1,2,3,4	(X5)	24	800	0.2514
Total			2 7 0	31 8 3	0. 9988

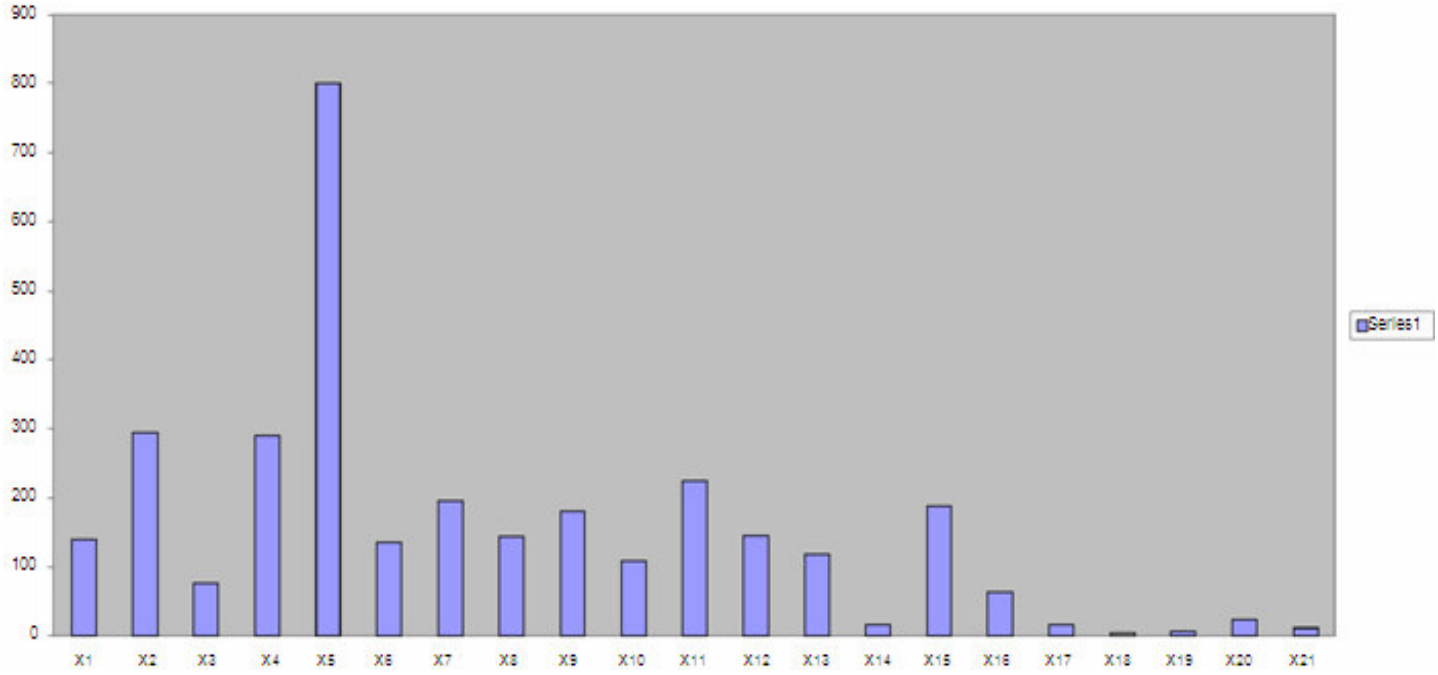

Figure 1. Frequency graph of nucleotide combinations in 3,183 helical turns of T4 phage DNA. The combinations are written in the order they appear in T4 phage DNA in the 5' to 3' direction.

in the particular message due to its source, or they may merely be common words and syllables which occur in any text in the language, such as the, and, tion, and the like in English language (Shannon, 1949). From this substitution of nucleotide combinations in T4 Phage with an English letter, over 300 words could be spelt out, phrases occurred occasionally, and a few statements could be made from the resulting text (Table 5). Reconstructions were often necessary to make sense ofthe statement.

DISCUSSION

Steganography, derived from the Greek, literally means "*covered writing*" (Johnson and Jajodia, 1998a). While cryptography scrambles a message so it cannot be understood, steganography hides a message so it cannot be seen (Johnson and Jajodia, 1998b).

DNA and molecular computing have ushered in a new era in computer science and information technology. The goal, however is not to compete with silicon based electronic computers, but rather to focus on areas of computation, such as combinatorics, where electronic computers are quite inefficient (Adleman, 1994). Moreover, as a information hiding medium, it is noteworthy that storing information in molecules of DNA allows for an information density of approximately 1 bit/nm^3, a dramatic improvement over existing storage media such as videotapes, which store information at a density of

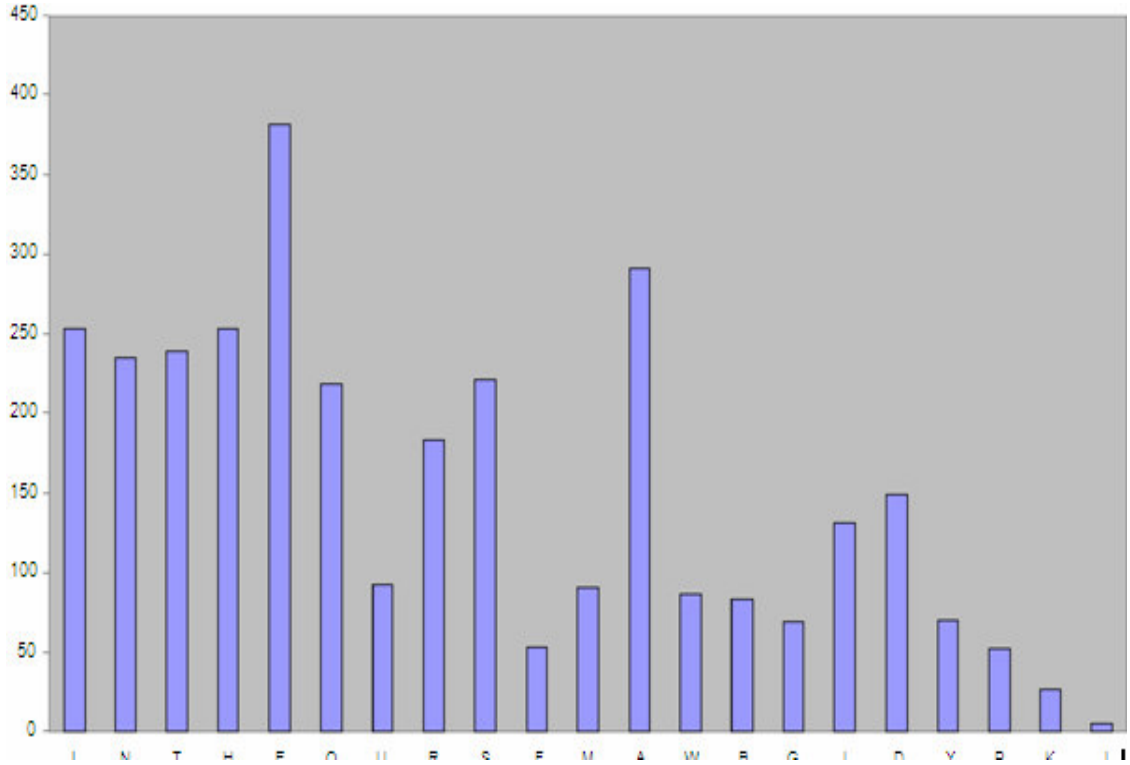

Figure 2. Frequency graph of 3,183 letters of chapter 5 of Wuthering heights. The letters are arranged in the order they appear in the text.

Table 3. Frequencies and probabilities of English letter out of 3,183 letters of an English text (Wuthering heights) in the case of the letters H and I which both occur 253 times, thereby having the same probabilities (the consonant is placed before the vowel).

S/N	Letter	Frequency	Probability
1	J	5	0.0016
2	K	26	0.0082
3	P	52	0.0163
4	F	53	0.0167
5	G	69	0.0217
6	Y	70	0.0220
7	B	83	0.0261
8	W	86	0.0270
9	M	91	0.0286
10	U	92	0.0289
11	L	131	0.0412
12	D	149	0.0468
13	R	183	0.0575
14	O	219	0.0688
15	S	221	0.0694
16	N	235	0.0738
17	T	239	0.0751
18	H	253	0.0795
19	I	253	0.0795
20	A	291	0.0914
21	E	382	0.1200
	Total	3183	1.0001

Table 4. A simple substitution table.

S/N	Nucleotide combination	English letter
1	0,0,5,5	J
2	0,0,2,8	K
3	0,1,1,8	P
4	1,1,1,7	F
5	0,0,3,7	G
6	0,0,4,6	Y
7	0,2,2,6	B
8	0,1,2,7	W
9	0,1,3,6	M
10	1,1,4,4	U
11	0,2,4,4	L
12	0,1,4,5	D
13	1,1,2,6	R
14	0,3,3,4	O
15	1,3,3,3	S
16	2,2,2,4	N
17	1,1,3,5	T
18	0,2,3,5	H
19	1,2,2,5	I
20	2,2,3,3	A
21	1,2,3,4	E

Table 5. Some words, phrases and statements found in the stegotext.

S/N	Words	Phrases	(Reconstructed) statement
1	IT	'BE HIM'	HO A SEAL
2	HE	'A SEA'	I TIRE
3	IS	'IT DIE'	WE EAT A TREE
4	ME	'LET TWO'	USE A SEA
5	US	'I READ'	HE IS WOE
6	THEIR	'AN HAT'	WOE HIT IT
7	DEN	'AS IS'	A NUN HE SEES
8	SUN	'SEE A SANE'	LET TIM'S TOE HEAL
9	MEN	'A RIDE'	HEAL US
10	ROB	'LORE DEN'	SO LET TWO LOAD
11	LIE	'USE LEG'	I READ IT SIR
12	SEED	'END IN'	IT FIT
13	AREA	'HER HAT'	BE RED ROSE
14	HEAL	'SELL BEAR'	YES HE ERRED HE SEES A NET
15	WEST	'BE OR'	
16	STAR	'I SUE'	
17	ELITE	'NUNS ARE'	
18	HEARD	'A REAL'	
19	TENET	'LET TWO'	
20	BERTH	'TEA SEED'	

approximately 1 bit/10^{12} nm^3 (Adleman, 1994). Currently, the only application of DNA in steganography is in the field of currency security, where special inks or materials with particular structure (such as fluorescent dyes or DNA) are used to write a hidden message on bank notes or other secure documents (Petitcolas et al., 1999). These materials provide a unique response to some particular excitation such as a reagent or laser light at a particular frequency.

SUMMARY AND CONCLUSION

The work is entirely hypothetical in nature, showing how DNA nucleotides mimic lettering in a natural language like the English language. While this work is purely hypothetical, the dramatic improvement in storage by DNA over conventional steganographic materials might have useful applications in the future. It is not impossible, that given the current outlay of designer molecular devices, the protocols of current molecular biology, genetic engineering and biotechnology might envision a future not unlike the skin implanted microchips, where information hiding could find a new terrain in microscopic cells, not to mention human subcutaneous cells. Presently, this work sheds new light and provides insight on the interesting way in which deoxyribonucleic acid nucleotides can be viewed as a hypothetical biochemical steganographic media, just as in standard computer science and engineering applications.

REFERENCES

Adleman LM (1994). Molecular computation of results to combinatorial problems. Science, 266:1021-1024.

Anderson RJ, Petitcolas FAP (1998). On the limits of steganography. IEEE J. Selected Areas Commun., Special Issue Copyright Privacy Protection, 16: 474-481.

Benenson Y, Adar R, Paz-Elizur T, Livneh Z, Shapiro E (2003). DNA molecule provides a computing machine with both data and fuel. Proc. Nat. Acad. Sci. USA, 100: 2191-2196.

Benenson Y, Paz-Elizur T, Adar R, Keinan E, Livneh Z, Shapiro E (2001). Programmable and autonomous computing machine made of biomolecules. Nature, 414: 430-434.

Bronte E (1965) Wuthering Heights. The Penguin English library: Middlesex.

Johnson NF, Jajodia S (1998a). Steganalysis: The analysis of hidden information. Proceedings of the IEEE Information Technology Conference, Syracuse, New York, USA.

Johnson NF, Jajodia S (1998b). Exploring steganography: seeing the unseen. IEEE Comput., 31: 26-34.

Margolin AA, Stojanovic MN (2005). Boolean Calculations made easy (for ribozymes). Nat. Biotechnol., 23: 1374-1376.

Miller ES, Kutter E, Mosig G, Arisaka F, Kunisawa T, Ruger W (2003). Bacteriophage T4. Microbiol. Mol. Biol. Rev., 67: 86-156.

Nelson DL, Cox MM (2000). Lehninger Principles of Biochemistry. New York: Worth Publishers.

Okamoto A, Tanaka K, Saito I (2004). DNA logic gates. J. Am. Chem. Soc., 126: 9458-9463.

Petitcolas FAP, Anderson RJ, Kuhn MG (1999). Information hiding: a survey. Proceed. IEEE, Special Issue Protection Multimedia Content, 87: 1062-1079.

Pirrung MC, Connors RV, Odenbaugh AL, Montague-Smith MP, Walcott NG, Tollet JJ (2000). The arrayed primer extension method for DNA microchip analysis. Molecular computation of satisfaction problems. J. Am. Chem. Soc.122: 1873-1882.

Quyang Q, Kaplan PD, Liu S, Libchaber A (1997). DNA solution of maximal clique problem. Science, 278:446-449.

Regev A, Shapiro E (2002). Cellular abstractions: cells as computation. Nature, 419: 343.

Sakamoto K, Gouzu H, Komiya K, Kiga D, Yokoyama S, Yokomori T, Hagiya M (2000). Molecular computation by DNA hairpin formation. Science, 288: 1223-1226.

Shannon CE (1949). Communication theory of secrecy systems. Bell Syst. Tech. J., 28: 656-715.

Stojanovic MN, Mitchell TE, Stefanovic DJ (2002). Deoxyribozyme-based Logic Gates. Am. Chem. Soc., 124:3555 – 3561.

Stojanovic MN, Stefanovic DJ (2003a). A deoxyribozyme based molecular automation. Nat. Biotechnol., 21: 1069-1074.

Stojanovic MN, Stefanovic DJ, (2003b). Deoxyribozyme based half-adder. J. Am. Chem. Soc., 125: 6673 – 6676.

Sutton WG, Rubin AD (2009). “Cryptography”, Microsoft® Encarta® DVD, Redmond WA: Microsoft Corporation.

Tacitus A (1990). How to survive under siege/Æneas the Tactician. Oxford: Clarendon ancient history series, Clarendon press, pp. 84-90, 183-193.

Weaver RF (2005). Molecular Biology, 3rd edition, McGraw-Hill.