GIAC
CERTIFICATIONS

Global Information Assurance Certification Paper

Copyright SANS Institute
Author Retains Full Rights

Interested in learning more?

Check out the list of upcoming events offering
"Security Essentials: Network, Endpoint, and Cloud (Security 401)"
at http://www.giac.org/registration/gsec

Donald Nixon
GSEC Assignment Version 1.3

DNA and DNA Computing in Security Practices – Is the Future in Our Genes?

Abstract

As modern encryption algorithms are broken, the world of information security looks in new directions to protect the data it transmits. The concept of using DNA computing in the fields of cryptography and steganography has been identified as a possible technology that may bring forward a new hope for unbreakable algorithms. Is the fledgling field of DNA computing the next cornerstone in the world of information security or is our time better spent following other paths for our data encryption algorithms of the future?

This paper will outline some of the basics of DNA and DNA computing and its use in the areas of cryptography, steganography and authentication.

Research has been performed in both cryptographic and steganographic situations with respect to DNA computing but researchers are still looking at much more theory than practicality. The constraints of its high tech lab requirements and computational limitations combined with the labour intensive extrapolation means, illustrate that the field of DNA computing is far from any kind of efficient use in today's security world. DNA authentication on the other hand, has exhibited great promise with real world examples already surfacing on the marketplace today.

DNA authentication practices will grow as the need for fool proof identification of individuals and items grows as well. The use of DNA computing on the other hand is far from a reality and the world of information security is better focused on other encryption technology methods for its future development.

Introduction

The world of encryption appears to be ever shrinking. Several years ago the thought of a 56 bit encryption technology seemed forever safe, but as mankinds' collective computing power and knowledge increases, the safety of the world's encryption methods seems to disappear equally as fast. Mathematicians and physicists attempt to improve on encryption methods while staying within the confines of the technologies available to us. Existing encryption algorithms such as RSA have not yet been compromised but many fear the day may come when even this bastion of encryption will fall by the way side as

have its predecessors.   There is hope for new encryption algorithms on the horizon utilizing mathematical principles such as Quantum Theory however the science of our very genetic makeup is also showing promise for the information security world.

The concepts of utilizing DNA computing in the field of data encryption and DNA authentication methods for thwarting the counterfeiting industry are subjects that have been surfacing in the media of late. How realistic are these concepts and is it feasible to see these technologies changing the security marketplace of today?

What is DNA?

Before delving into the principles of DNA computing, we must have a basic understanding of what DNA actually is. All organisms on this planet are made of the same type of genetic blueprint which bind us together. The way in which that blueprint is coded is the deciding factor as to whether you will be bald, have a bulbous nose, male, female or even whether you will be a human or an oak tree.
Within the cells of any organism is a substance called Deoxyribonucleic Acid (DNA) which is a double-stranded helix of nucleotides which carries the genetic information of a cell. This information is the code used within cells to form proteins and is the building block upon which life is formed.

Strands of DNA are long polymers of millions of linked nucleotides. These nucleotides consist of one of four nitrogen bases, a five carbon sugar and a phosphate group. The nucleotides that make up these polymers are named after the nitrogen base that it consists of; Adenine (A), Cytosine (C), Guanine (G) and Thymine (T). These nucleotides will only combine in such a way that C always pairs with G and T always pairs with A. The two strands of a DNA molecule are antiparallel where each strand runs in an opposite direction. Figure 1 illustrates two strands of DNA and the bonding priciples of of the 4 types of nucleotides and the Figure 2 illustrates the double helix shape of DNA.
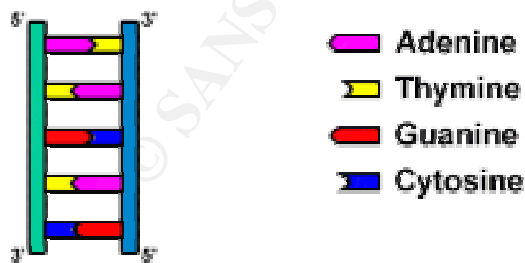

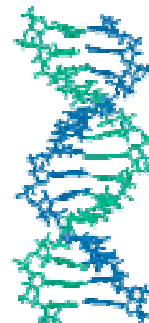Fig 1 – Graphical representation of inherent bonding properties of DNA [11]

Fig 2 – Illustration of double helix shape of DNA. [11]

The combination of these 4 nucleotides in the estimated million long polymer strands can result in billions of combinations within a single DNA double-helix. These massive amount of combinations allows for the multitude of differences between every living thing on the planet from the large scale (mammal vs. plant), to the small (blue eyes vs. green eyes).

With the advances in DNA research in projects such as the Human Genome project (a research effort to characterize the genomes of human and selected model organisms through complete mapping and sequencing of their DNA [17]) and a host of others, the mystery of DNA and its construction is slowly being unraveled through mathematical means. Distinct formulae and patterns have emerged that may have implications well beyond those found in the fields of genetics.

What does all this chemistry and biology have to do with security you might ask? To answer that question we must first look at how biological science can be applied to mathematical computation in a field known as DNA computing.

Basics and Origins of DNA Computing

DNA computing or molecular computing are terms used to describe utilizing the inherent combinational properties of DNA for massively parallel computation. The idea is that with an appropriate setup and enough DNA, one can potentially solve huge mathematical problems by parallel search. Basically this means that you can attempt every solution to a given problem until you came across the right one through random calculation. Utilizing DNA for this type of computation can be much faster than utilizing a conventional computer, for which massive parallelism would require large amounts of hardware, not simply more DNA. [10]

Leonard Adleman, a computer scientist at the University of Southern California was the first to pose the theory that the makeup of DNA and it's multitude of possible combining nucleotides could have application in brute force computational search techniques. Adleman is also known as the 'A' in the RSA algorithm - an algorithm that in some circles has become the de facto standard for industrial-strength encryption of data sent over the Internet.

In early 1994, Adleman put his theory of DNA computing to the test on a problem called the Hamiltonian Path problem or sometimes referred to as the Traveling Salesman Problem. The 'salesman' in this problem has a map of several cities that he must visit to sell his wares where these cities have only one-way streets between some but not all of them. The crux of the problem is that the salesman must find a route to travel that passes through each city (A through G) exactly once, with a designated beginning and end. (Fig. 3) The salesman wants to make efficient use of his time and does not want to backtrack or double back on a path he has already taken previously.
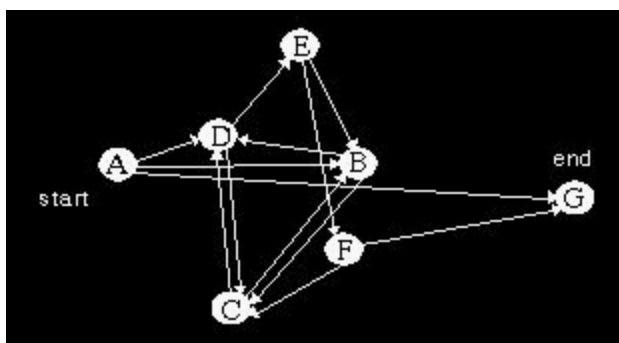
Fig. 3 – Basic outline of 'Traveling Salesman' Problem representing the 7 cities and one way streets between them.

This type of problem is known as a non-deterministic polynomial time problem (NP). The idea of guessing the right answer to a problem, or checking all possible problems in parallel to determine which is correct, is called nondeterminism. An algorithm that works in this manner is called a nondeterministic algorithm, and any problem with an algorithm that runs on a non-deterministic machine in polynomial time is called a non-deterministic polynomial time problem.

The NP problem was chosen for Adleman's DNA computing test as it is a type of problem that is difficult for conventional computers to solve. Conventional computers are better suited for deterministic computation permitting at most one next move at any step in a computation. The inherent parallel computing ability of DNA combination however is perfectly suited for NP problem solving.

Adleman, using a basic 7 city, 13 street model for the Traveling Salesman Problem, created randomly sequenced DNA strands 20 bases long to chemically represent each city and a complementary 20 base strand that overlaps each city's strand halfway to represent each street (Fig. 4). This representation allowed each multi-city tour to become a piece of double stranded DNA with the cities linked in some order by the streets.

Fig 4. – Representation of 20 base DNA strand representing a city showing the bonding tendencies of nucleotides to DNA strands representing pathways between the cities.

By placing a few grams of every DNA city and street in a test tube and allowing the natural bonding tendencies of the DNA building blocks to occur, the DNA bonding created over 10^9 answers in less than one second. [5] Of course, not all of those answers

that came about in that one second were right answers as Adleman only needed to keep those paths that exhibited the following properties:

1. The path must start at city A and end at city G.
2. Of those paths, the correct paths must pass through all 7 cities at least once.
3. The final path(s) must contain each city in turn.

Now the initial calculation took approximately one second but Adleman's extrapolation was performed over a period of a week. The 'correct' answer was determined by filtering the strands of DNA according to their end-bases to determine which strands begin from city A and end in city G and discarding those that did not. The remaining strands were then measured through electrophoreic techniques to determine if the path they represent has passed through all 7 cities. Finally the resulting sets of strands were examined individually to determine if they contained each city in turn. That strand or strands that remained was then determined to be the answer or equivalent answers.

Adleman found his one true path for the 'Salesman' in his problem and the possible future of DNA computing opened up in front of him. Granted, these initial experiments were performed on a small number of cities and the same answer can be quickly determined in about an hour using a pencil, paper and a sharp mind, but the ability to solve problems with larger numbers of cities and paths using the same techniques was immediately feasible.

To Adleman, the following advantages of DNA computing became evident;

Speed – Conventional computers can perform approximately 100 MIPS (millions of instruction per second). Combining DNA strands as demonstrated by Adleman, made computations equivalent to 10^9 or better, arguably over 100 times faster than the fastest computer. [5] The inherent parallelism of DNA computing was staggering.

Minimal Storage Requirements – DNA stores memory at a density of about 1 bit per cubic nanometer where conventional storage media requires 10^{12} cubic nanometers to store 1 bit. [5] In essence, mankinds collective knowledge could theoretically be stored in a small bucket of DNA solution.

Minimal Power Requirements - There is no power required for DNA computing while the computation is taking place. The chemical bonds that are the building blocks of DNA happen without any outside power source. There is no comparison to the power requirements of conventional computers.

There have been other researchers since Adlemans work that have demonstrated similar possibilities of DNA computing. For example a group of researchers at Princeton in early 2000 demonstrated an RNA computer similar to Adleman's which had the ability to solve a chess problem involving how many ways there are to place knights on a chess board so that none can take the others.

Adleman instantly envisioned the use of DNA computing for any type of computational problems that require massive amounts of parallel computing. As his background stemmed from computer encryption, he particularly envisioned DNA computing in helping create and decipher algorithms in the field of cryptography. The possibility existed of the very genetic makeup of an individual being used in the encryption/decryption of data from/to that person. The possibility was also seen that the DNA of an individual will give them the 'who you are' portion of the 'who you are', 'what you know', 'what you have' aspects of security authentication.

There has been much speculation of the use of this type of technology for cryptographic and steganographic means that would take advantage of the parallel computation possibilities available with DNA computing. Are these real possibilities for the security industry or are the processes involved in its implementation too difficult to envision in the immediate future? Next we will examine DNA computing within these security applications to determine the likelihood of DNA being used in their advancement.

DNA Cryptography

DNA cryptography has been bantered about much in the media as of late but whether or not this technology is appropriate for the future is debatable. There has been a distinct lack of hard evidence put forward to illustrate whether the technology is even feasible, much less appropriate in the foreseeable future.

Ashish Gehani, Thomas LaBean and John Reif of Duke University have published a paper entitled 'DNA-based Cryptography' which puts an argument forward that the high level computational ability and incredibly compact information storage media of DNA computing has the possibility of DNA based cryptography based on one time pads. They argue that current practical applications of cryptographic systems based on one-time pads is limited to the confines of conventional electronic media whereas as small amount of DNA can suffice for a huge one time pad for use in public key infrastructure (PKI). [1]

To put this into terms of the common Alice and Bob description of secure data transmission and reception, they are basing their argument of DNA cryptography on Bob providing Alice his public key, and Alice will use it to send an encrypted message to him. The potential eavesdropper, Eve, will have an incredible amount of work to perform to attempt decryption of their transmission than either Alice or Bob.
Public key encryption splits the key up into a public key for encryption and a secret key for decryption. It's not possible to determine the secret key from the public key. Bob generates a pair of keys and tells everyone his public key, while only he knows his secret key. Anyone can use Bob's public key to send him an encrypted message, but only Bob knows the secret key to decrypt it. This scheme allows Alice and Bob to communicate in secret without having to physically meet as in symmetric encryption methods. [15]
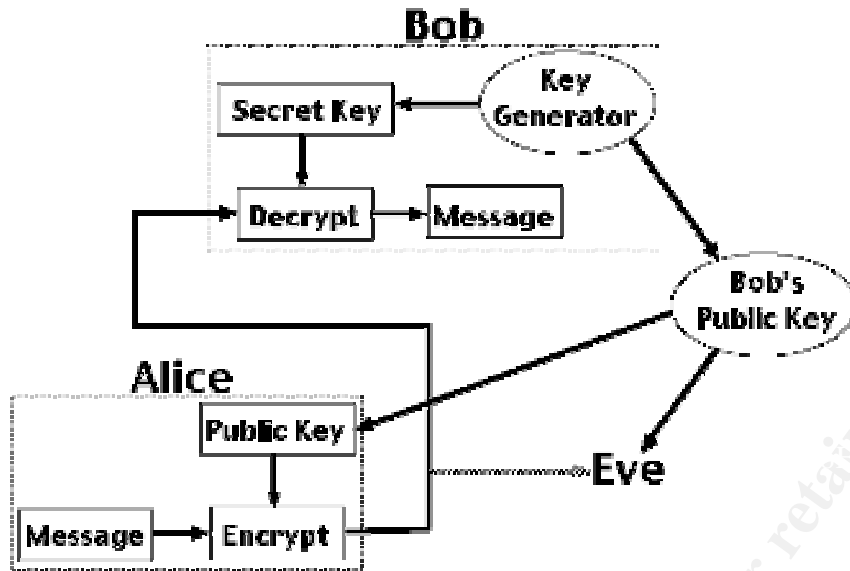
Fig 5. Public Key Encryption illustrated. [16]

Injecting DNA cryptography into the common PKI scenario, the researchers from Duke argue that we have the ability to follow the same inherent pattern of PKI but using the inherent massively parallel computing properties of DNA bonding to perform the encryption and decryption of the public and private keys. In essence, the encryption algorithm used in the transaction can now be much more complex than that in use by conventional encryption methods.

It can easily be argued that DNA computing is just classical computing, albeit highly parallelized; thus with a large enough key, one should be able to thwart any DNA computer that can be built. [10]   This puts the idea of this form of DNA computing at great risk in the field of cryptography. As well, the obstacles of utilizing this kind of technology outside of a lab are extremely high. There is a paramount need for a lab environment for both the combination of DNA strands and the extrapolation of the 'answers' that those combinations will provide.

DNA Cryptography has yet to be proven on anything but paper as of yet but there has been some examples set forward in the field of DNA steganography that are worth note. First we will examine some basics of Steganography and then we will delve into the marriage of DNA and steganography.

Origins of Steganography

Steganography is a variety of encryption that completely hides text or graphics, usually unencrypted, within other text or graphics that are electronically transmitted. [18]   The

science of steganography has been with us since early Grecian times and its definition has come to envelop much more technology than the ancient Greeks ever imagined possible.

The term steganography derives from the Greek words **steganos** meaning hidden and **graphein** meaning to write. One of the early Grecian methods of steganography was to shave the head of a messenger, tattoo the message to be hidden on the messengers head and then allowing the hair to grow back before sending him on his way. The hidden message could not then be uncovered until the messengers head was shaved bald once again.

Throughout our history there have been many other forms of steganography used to hide messages such as the use of null ciphers, invisible inks and others. In World War II for example, German cryptographers devised a method of using microdots to conceal messages within messages themselves. Photographs containing a message were shrunken to miniature proportions and placed on an inconspicuous piece of correspondence. The first detection of this microdot technology was actually found masquerading as a period at the end of a sentence on an envelope carried by a German soldier. The microdot was not actually encrypted or even really hidden at all – just inconspicuously small enough to avoid detection.

More recently, computer technology and the Internet have provided a medium for steganography that has been unseen in the past. The ability to transfer text and images is now instantaneous and accessible by individuals virtually everywhere on the planet. It has been reported that the Al Queda network of terrorists may have used steganographic means to hide their communications in organizing the September 11th attacks on the United States of America.

Readily available software applications such as the freeware application JPHide and JPSeek will encrypt messages with the common JPG format of graphic files. Other applications give the user the ability to hide data within other graphic formats such as GIF or BMP and audio formats such as MP3. Messages can now be hidden in the inconspicuous advertising banners of web pages and the music files we listen to.

Although older forms of steganography are unencrypted, much of today's steganography found in the electronic medium is in an encrypted format like that found in programs such as JPHide and JPSeek. Much like the world of data transmission, the steganographic world is on the lookout for the encryption methods that cannot be broken. Can DNA steganography provide that unbreakable encryption medium?

DNA Steganography

Experiments in DNA Steganography have been conducted by Carter Bancroft and his team at the Mt. Sinai School of Medicine to encrypt hidden messages within microdots. Bancroft using the microdot methodology utilized in message hiding during World War

II data transmission has done something similar this time using DNA encoded microdots. (Fig. 6)

The principles used in this experiment used a simple code to convert the letters of the alphabet into combinations of the four bases which make up DNA and create a strand of DNA based on that code. A piece of DNA spelling out the message to be hidden is synthetically created which contains the secret encrypted message in the middle plus short marker sequences at the ends of the message. The encoded piece of DNA is then placed into a normal piece of human DNA which is then mixed with DNA strands of similar length. The mixture is then dried on to paper that can be cutup into microdots with each dot containing billions of strands of DNA. Not only is the microdot difficult to detect on the plain message medium but only one strand of those billions within the microdot contains the message. [6]

The key to decrypting the message lies in knowing which markers on each end of the DNA are the correct ones which mean there must be some sort of shared secret that is transmitted previously for this type of transmission to work successfully. Once the strand is determined via identifying the markers, the recipient uses polymerase chain reaction to multiply only the DNA which contains the message and applies the simple code to finally decode the true message. [2] Utilizing these methods, Bancroft and his team were successfully able to encode and decode the famous message 'June 6 Invasion: Normandy' within a microdot placed in the full stops on a posted typed letter.
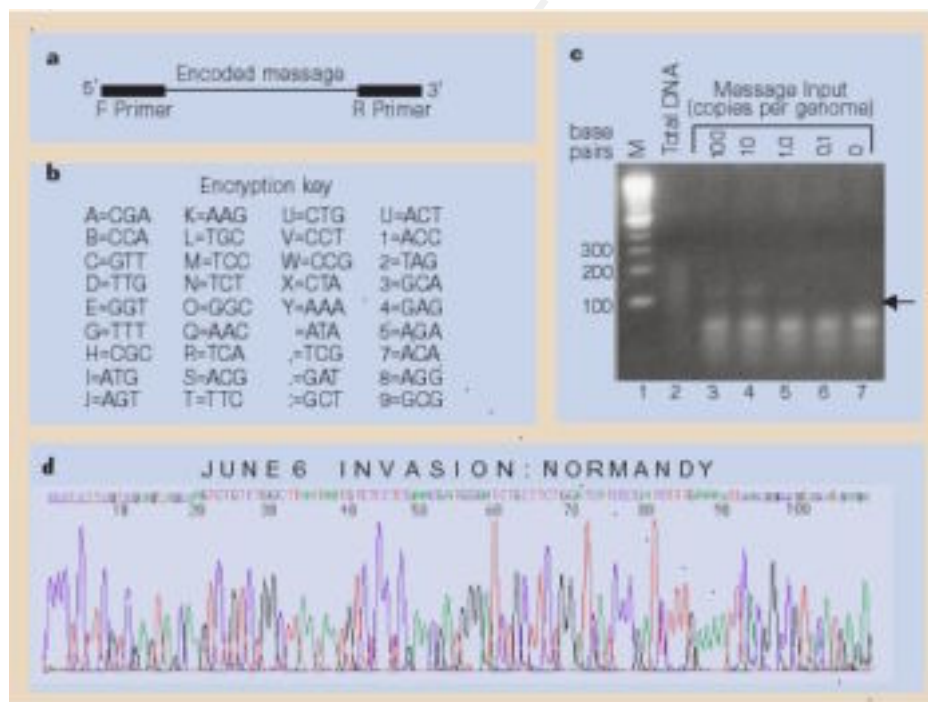

Fig 6. DNA Steganography. a, Structure of secret message DNA strand illustrating marker sequences. b, key used to encode message in DNA. c, Gel analysis of DNA strand. d, Sequence of cloned product of PCR amplification and resulting encoded message. [8]

Perhaps this all appears a bit far fetched at first and the skeptics state that the same problems with the DNA cryptography are evident in DNA steganography as well. The 'test tube' environment used in this type of steganography is far from practical for everyday use. The DNA microdot team does see this technology having applications in another field however – that of authentication. With the amount of plant and animal genetic engineering that is taking place today and will continue to do so in the future, this methodology would allow engineers to place DNA authentication stamps within organisms they are working with to easily detect counterfeits or copyright infringements.

DNA Authentication

It is worth mentioning that DNA authentication is currently at work in the marketplace today albeit not in the genetic engineering form envisioned by Bancroft and his team. Forms of DNA authentication have already been used for such items as the official clothing from the Sydney Olympic Games, sports collectibles and limited edition art markets such as original animation cells distributed by the Hanna Barbara group of artists.

In the case of the clothing used in the Sydney Olympic Games, a Canadian company named DNA Technologies was able to showcase its DNA-tagging abilities on the world stage in the summer of 2000. All Olympic merchandise from shirts and hats to pins and coffee mugs were tagged with special ink that contained DNA taken from an unnamed Australian athlete. DNA was taken via saliva samples from the athlete and mixed into existing ink compounds which was in turn used in the regular merchandise manufacturing process. A hand held scanner is then used to scan the inked area of the clothing to determine if a piece of merchandise is authentic or not. As it is estimated that the human genome is roughly 3 billion base pairs in size, and the samples taken were from a random athlete from a Olympic team of hundreds, the possibility of counterfeiting this merchandise is difficult to say the least. [4]   For the Sydney games, DNA inks were applied too nearly 50 million items at a cost of about five cents each, including licensing, databasing, and back-end support.

There are possibilities of this type of technology to be used in the arenas of currency and other such brandable items where existing authentication methods such as holograms are proving ineffective and costly. DNA-tagging is much cheaper in comparison and ultimately more difficult to thwart.

Conclusion

The field of DNA computing is still in its infancy and the applications for this technology are still not fully understood. The world of information security is always on the lookout for unbreakable encryption to protect the data that we transmit but it appears that every

encryption technology meets its endgame as the computing technology of our world evolves. It appears we are involved in a paradox where the best encryption technology of the day is only as good as the computing power that it is tested upon and the practicality of its application. Is DNA computing viable – perhaps, but the obstacles that face the field such as the extrapolation and practical computational environments required are daunting. DNA authentication methods on the other hand have shown great promise in the marketplace of today and it is hoped that its applications will continue to expand.

The beauty of both these DNA research trends is found in the possibility of mankinds' utilization of its very life building blocks to solve its most difficult problems. In any case, we will not be tossing out those PC's for test tubes of DNA anytime soon and the use of DNA computing with a greater security focus other than in merchandise authentication methods is a long way off.

References:

1. Gehani, Ashish. La Bean, Thomas H. Reif, John H. "DNA-Based Cryptography". Department of Computer Science, Duke University. June 1999, http://www.cs.duke.edu/~reif/paper/DNAcrypt/crypt.pdf
2. "Sci/Tech DNA hides spy message". BBC.co.uk. June 10, 1999. http://news.bbc.co.uk/hi/english/sci/tech/newsid_365000/365183.stm
3. Pelletier, Oliver. "Algorithmic Self-Assembly of DNA Tiles and its Application of Cryptanalysis". October 2, 20001. http://xxx.lanl.gov/abs/cs.CR/0110009
4. Taggart, Stewart. "Call it the SyDNA Olympics". Wired.com. March 7, 2000. http://www.wired.com/news/technology/0,1282,34774,00.html
5. Gupta, Gaurav. Mehra, Nipun. Chakraverty, Shumpa. "DNA Computing". The Indian Programmer. June 12, 2001. http://www.theindianprogrammer.com/technology/dna_computing.htm
6. Peterson, Ivars. "Hiding in DNA". Science News Online. April 8, 2000. http://www.sciencenews.org/20000408/mathtrek.asp
7. Blahere, Kristina. "DNA Computing". CNET. April 26, 2000. http://www.cnet.com/techtrends/0-1544316-7-1727815.html?tag=st.sr.1544316-7-1727814.arrow.1544316-7-1727815
8. Taylor Clelland, Catherine. Risca, Viviana. Bancroft, Carter. "Hiding Messages in DNA Micodots". Nature Magazine Vol 399. June 10, 1999. http://www.cs.memphis.edu/~garzonm/pub_old/datsec/dnastegano.pdf
9. "Algorithms and Theory of Computational Handbook , CRC Press LLC , 99" (provided by the National Institute of Standards and Technology). February 2, 2002. http://www.nist.gov/dads/

10. "Frequently Asked Questions About Today's Cryptography 4.1 - Section 7.19 What is DNA Computing". RSA Laboratories. http://www.rsasecurity.com/rsalabs/faq/7-19.html

11. Team # 18617. "DNA – Prelude to the Symphony of Life". ThinkQuest.org. http://library.thinkquest.org/18617/data/types/dna.html?tqskip1=1&tqtime=0212

12. Johnson, R. Colin. "RNA computer clears 10-bit hurdle". EE Times. February 1, 2000. http://www.eetimes.com/story/OEG20000201S0007

13. Friedman, Yali. "DNA Based Computers". http://dna2z.com/dnacpu/dna2.html

14. Adleman, Leonard. "Molecular computation of solutions to combinatorial problems". Science 266, 1021-1024. November 11, 1994

15. "PKI Infrastructure". Treasury Board of Canada. October 4, 2001. http://www.cio-dpi.gc.ca/pki-icp/beginners/whatisapki/whatisapki_e.asp

16. Johnson, Paul. "Introduction to PKI". January 27, 2001. http://pajhome.org.uk/crypt/rsa/intro.html

17. "The Human Genome Project". February 9, 2000. http://www.nhgri.nih.gov/HGP/#What

18. "Telecom Glossary 2000". August 3, 2001. http://www.its.bldrdoc.gov/projects/t1glossary2000/_steganography.html

19. Gaudin, Sharon. "The terrorist network". Network World. November 26, 2001. http://www.nwfusion.com/research/2001/1126featside4.html

20. Taggart, Stewart. "DNA's Olympic Trial". Business 2.0. August 2000. http://www.business2.com/articles/mag/0,1640,13953,FF.html

21. McCullagh, Declan. "Bin Laden: Steganography Master?". Wired News. February 7, 2001. http://www.wired.com/news/politics/0,1283,41658,00.html

22. Johnson, Neil F. "Steganography". http://www.jjtc.com/stegdoc/index2.html