Network Security via biometric recognition of patterns of gene expression

Harry C Shaw
Telecommunication Networks & Technology Branch
NASA/Goddard Space Flight Center
Greenbelt, MD
Harry.c.shaw@nasa.gov

***Abstract*— Molecular biology provides the ability to implement forms of information and network security completely outside the bounds of legacy security protocols and algorithms. This paper addresses an approach which instantiates the power of gene expression for security. Molecular biology provides a rich source of gene expression and regulation mechanisms, which can be adopted to use in the information and electronic communication domains. Conventional security protocols are becoming increasingly vulnerable due to more intensive, highly capable attacks on the underlying mathematics of cryptography. Security protocols are being undermined by social engineering and substandard implementations by IT organizations. Molecular biology can provide countermeasures to these weak points with the current security approaches. Future advances in instruments for analyzing assays will also enable this protocol to advance from one of cryptographic algorithms to an integrated system of cryptographic algorithms and real-time expression and assay of gene expression products.**

Keywords—network security; gene expression; transcription factors; translation;

I. INTRODUCTION

Network security is a vital component of the design of any network. There are five main requirements to be addressed in developing a secure network: Authentication, confidentiality, data integrity, non-repudiation, and access control. *In vivo*, biomolecular cellular systems of gene expression authenticate themselves through various means such as transcription factors and promoter sequences. These factors also enforce access control. They have means of retaining confidentiality of the meaning of genome sequences through processes such as control of protein expression. They are capable of establishing data integrity and non-repudiation through transcriptional and translational controls.

A suite of genomics and proteomics based authentication and confidentiality protocols are being developed to augment traditional network security approaches with concepts from molecular biology via the regulation of gene expression. These protocols are agnostic to their implementation and can be incorporated into any existing network security protocol (Secure http, SSL, TLS, IPSec, etc.) or any future network security strategy. The protocols can be implemented for implementing web-based security strategies, digital signatures, digital rights management, and general purpose encryption for data in motion or data at rest.

Initial implementations will utilize the cryptographic algorithms described below. Future implementations will take advantage of the ability to access patterns of gene expression *in vivo*. Knowledge of gene expression products (proteins and non-coding RNA) can be achieved without time consuming sequence assays. It can be done via techniques such as fluorescent labeling. In fact the use of fluorescence is one method of implementing a protein expression security protocol.

These protocols will provide new challenges for network attackers by forcing them to work in both the information security domain and the molecular biology domain. Although no security strategy is without vulnerabilities, the intent of this work is to present a completely new set of problems for network attackers which will result in higher network and information security.

II. IMPLEMENTATION OF A MOLECULAR BIOLOGICALLY BASED SECURITY PROTOCOL

Implementation of this approach utilizes the following concepts:

- A network security concept of operations based upon the processes of gene expression.
- A methodology for taking the processes of gene expression and converting them into cryptographic protocols (ciphergenes become analogs for biological genes, cipherproteins become analogs for biological proteins)
- Specific coding models for the cryptographic protocols
- Cryptographically hard sources for the patterns of gene expression called ciphercolonies. (ciphercolonies become analogs for colonies of living organisms but can also contain algorithms substituting for living organisms)
- Ultimate merging of security protocols with *in vivo* and *in vitro* realizations of ciphercolonies. Networks of ciphercolonies capable of signaling and responding to patterns of gene expression within a network and authenticating members of the network.

Security is based upon the translation of plaintext messages to and from a computationally large set of genomic and proteomic messages. Instead of relying solely on the four nucleotides and 20 amino acids as a code base, messages are generated from genomic and proteomic sequences that include the informational representation of the regulatory elements of transcription and translational networks. A single sequence can generate many messages depending upon the transcription and translation instructions. Modules can be developed and implemented in large or small systems, firewalls, routers, switches and other devices. Ultimately, the goal is a network of biologically enabled nodes and Certificate Authorities who can establish trust via evolving patterns of gene expression. Like any security protocol, the most efficient implementation requires accommodating its features early in the design phase. To accomplish that goal, designers must understand the key concepts and future IT security departments will require molecular biologists and biochemists on their staffs.

A. *Gene Expression Nomenclature*

The processes of transcription and translation are coded into a protocol that assigns each functional region of a relevant molecule a code derived using the Method of Types from information theory [1]. For the process of transcription the protocol focuses on the interaction between regulatory sequences in the gene, transcription factors that must bind to the gene, and the enzyme that binds to the entire complex to perform transcription from DNA to messenger RNA. The analogy in the protocol is series of codes from the Method of Types. There can be a code for each of the regulatory sequences, transcription factors, and enzyme. The interaction between codes is defined by a joint probability matrix for the intersection of 2 or more codes. The intersection of the codes is analogous to a binding interaction. In the encryption step, all of those interactions are specified. In the decryption step, the receiver decodes those interactions using the same joint probability distributions as the sender. An example is shown in fig. 1 [2]. There are six general transcription factor proteins (TFIIA, TFIIB, TFIID, TFIIE, TFIIF, TFIIH) shown. There are five regulatory sequences on the gene (BRE, TATA, INR, MTE, DPE). A set of interactions are defined for this transcriptional complex a subset of which is shown in table 1.

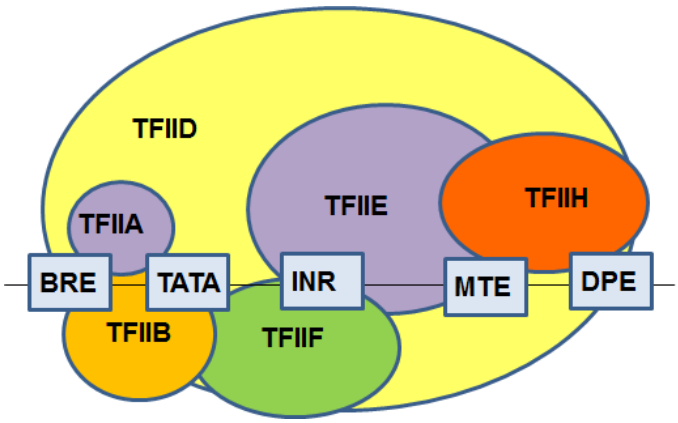

Fig. 1. Coding the transcriptional complex

TABLE I. SAMPLE OF EVENT JOINT PROBABILITIES FOR TRANSCRIPTIONAL COMPLEX

Event (Protein-DNA)
TFIIA ∩BRE
TFIIA ∩ TATA
TFIIB∩BRE
TFIIB∩TATA
TFIIE∩INR
TFIIE∩MTE
TFIIA ∩ TATA

B. *Application of the Method of Types*

If we define the relationship between protein transcription factor and regulatory sequence in terms of jointly typical sets of the two sequences, then different levels of homology can be required in different authentication or confidentiality scenarios. As an example: Let $\Gamma = \{1, 2, 3, 4, 5\}$, a 5-tuple alphabet for regulatory sequences with Type Γ_g, 10 digits in length, consisting of the distribution shown in table 2.

TABLE II. EXAMPLE OF A TYPE Γ DEFINTION

$P_{g1}=0.2$
$P_{g2}=0.4$
$P_{g3}=0.1$
$P_{g4}=0.1$
$P_{g5}=0.2$

The type class of Γ_g consists of all sequences within Γ with the same statistical distribution, as shown in (1).

$$T(\Gamma_g)=\{1122223455,\ldots,554322221\} \quad (1)$$

Codes from $T(\Gamma_g)$ can be defined for different transcription factors of the family TFII. Then the binding criteria is defined as the mutual information between sets of codes using joint probability distributions. This is shown as the intersection between the codes. The concept includes all interactions between protein codes and nucleotide codes. Codes without interactions have joint probabilities that yield the null set.

The power of the Method of Types is that it allows for a certain amount of statistical variability in the interaction of the codes which mimics the variability of cellular interactions. Table 1 provides an example of the interpretation of binding of the general transcription factor proteins to a ciphergene gene undergoing encryption or decryption. In fig.1 all of the transcription factor proteins must bind to the correct gene regulatory sequences and other transcription factor proteins to permit RNA Polymerase II to bind to the entire complex and perform transcription. From a coding perspective, all of the codes for these factors have a defined probabilities of code intersection that signal successful authentication and permit subsequent operations of encryption or decryption to occur.

Fig. 2 depicts a mobile, ad-hoc network (MANET) of six users challenging an incoming user. In this case, the challenge is to provide a coded response representing the network diagram shown in fig. 1. In this case the responder with ciphergene Z must know each cipherprotein-cipherprotein interaction and each cipherprotein-ciphergene interaction and only respond with the valid interactions. Then each of the six existing members decodes the response for their specific cipherprotein-cipherprotein and ciphergene interactions to determine admission to the network. In this example, using fig.1 as a guide, User A codes the cipherprotein-cipherprotein code intersection for TFIIA∩TFIIB and the cipherprotein-ciphergene code intersection for TFIIA∩BRE and TFIAA∩TATA. Those codes are sent as the MANET challenge to a user who supposedly possesses ciphergene Z, who responds to User A with the individual code types for TFIIA, TFIIB, BRE and TATA. All other users perform the same function for their respective cipherprotein-ciphergene and cipherprotein-cipherprotein codes. Successful completion by holder of ciphergene Z allows admission to the MANET.

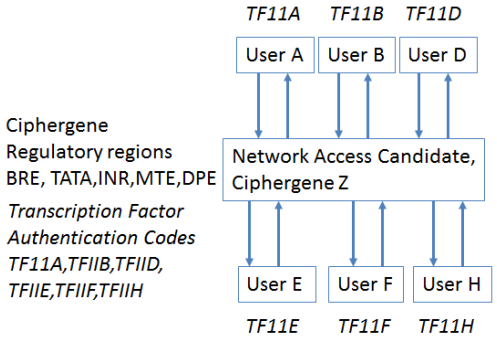

Fig. 2. MANET Authentication with general transcription factor and gene regulatory sequence coding

C. *Design of the Genomic and Proteomic Protocol suite*

1. *Coding of sequences as objects.*

An object is defined as a genomic or proteomic sequence. It could be sequence defined at the nucleotide base level (e.g. AGGCT…) the codon level, (AAG, TTA, CGC, …), transcription factor/ binding site (SP1, CCAT, AP2,…), protein transcription factor (TFIIA, TFIIB,…) , and so forth.

Each object is drawn from the elements in a dictionary set associated with that object, for example:

- Nucleotides: *{A, T, C, G, U, I, MeC, X, H} A, G, C,* and *T* represent the main DNA bases adenine, guanine, cytosine, and thymine. *MeC* represents 5-Methylcytosine, an epigenetic marker, *H* represents hypoxanthine, and *X* represents xanthine. *H* and *X* are mutagenic deaminations of DNA bases that occasionally occur in gene sequences.
- DNA Codons: *{ATT, ATC, …, , TGA}*
- Transcription factors: *TF ={TFII, TBP, …}*

Associated with each dictionary is a set of class elements that describe the function of an element in a given sequence. For the set of nucleotides, *N*, the classes might be:

- Regulatory elements: *{Promoter, Upstream Activator, TATA, Exon, Intron, …}*

2. *High-level implementation summary*

First, there is a level 1 process by which DNA text is mapped into the structure of a gene complete with introns, exons, regulatory regions, etc., to create a ciphergene. The purpose of this coding from a security perspective is that a single sequence of letters from a small alphabet can be used to represent a large set of permutations of message combinations. Multiple messages can be encoded into a single ciphergene.

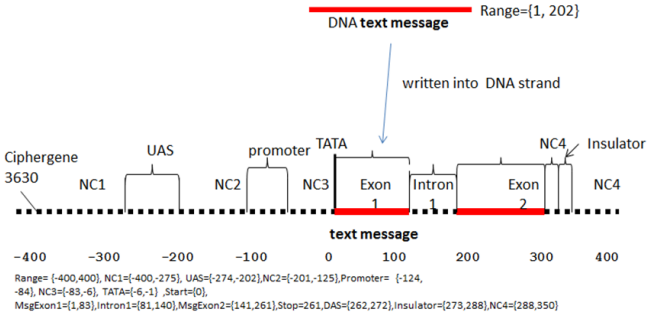

Fig. 3.Biological gene structure with a pre-coding instantiation of the DNA text message

The decoding of such messages represents an *np*-hard problem for attackers. Fig. 3 provides a simple example of how the encoding process begins. In a real application the message could be spread across the control regions and different messages could be encoded simultaneously across different regions.

Level 1 can be used as an off-ramp and transmitted to the receiver or proceed to level 2. Level 2 processes take the ciphergene code through a series of protein transcription factor code operations that combine with their counterpart regulatory codes on the ciphergene to produce a new coded sequence that represents a coded transcriptional complex.

Similarly, level 2 can be used as an off-ramp and transmitted to the receiver or proceed to level 3. Level 3 processes are a series of operations that takes the coded transcriptional complex, through a series of operations using protein and RNA polymerase codes resulting in a basal transcriptional complex code. The basal transcriptional complex code is processed by algorithms and maps the code into a messenger RNA code, called the cipher-mRNA code. The cipher-mRNA now consists only of codons of the original DNA text message and is translated into a protein code, called the cipherprotein.

The output of level 3 is the cipherprotein code that is transmitted from the sender to the receiver. The receiver applies the decryption keys to recover the cipher-mRNA and then perform all subsequent steps to reach level 2, level 1, and decoding to produce the plaintext. The protocol uses a series of encryption and decryption matrices, a series of encryption keys derived from pre-shared secret genomic sequences and additional tools from information theory to accomplish coding,

encrypting, decoding and decrypting the successive levels of data [3].

III. ESTABLISHING AND UTILIZING PATTERNS OF GENE EXPRESSION

A. *Utilization of gene expression pathways*

A pattern of gene expression is created by an organism going through the processes of DNA transcription and RNA translation across the many genes within the genome. Genes are always expressed within the context of overall cellular requirements. Thus, genes are expressed in response to stimuli indicating a need for expression. Not all genes are expressed all the time. This introduces a temporal element into the security. An attacker might know which organisms are the basis of the security system, but not necessarily have knowledge of when genes are expressed in response to given stimuli. The more complex the organism, the greater the diversity of the pattern of gene expression which will contribute to diffusion and confusion of the codes generated by the protocols. Mutation processes increase the security provided these protocols because cryptanalysis of the text cannot be fully relied upon the retrieve the plaintext. And mutations do not always modify the sequence of an expressed protein.

The factors of mutation and temporal gene expression can be used to increase the overall security of a system.

B. *Implementation of the protocols in living systems.*

These protocols are designed for eventual implementation in cellular organisms and colonies, when the technology is ready. By keeping the coding rules consistent with mechanics of gene expression, it is expected that there will eventually be instruments that read the cryptographic instructions and produce gene expression products *in vivo*. Networks will interface with each other over time, such that, as a minimum a level of phenotypic recognition can be established between networks. Mutation and variation play an important role in the security aspects. A Man-In-The-Middle attacker that has not been in constant contact with the colonies will be unable to match the patterns of gene expression that have evolved over time between 2 or more networks that recognize each other. This leads to self-authenticating network interactions as shown in fig. 4.

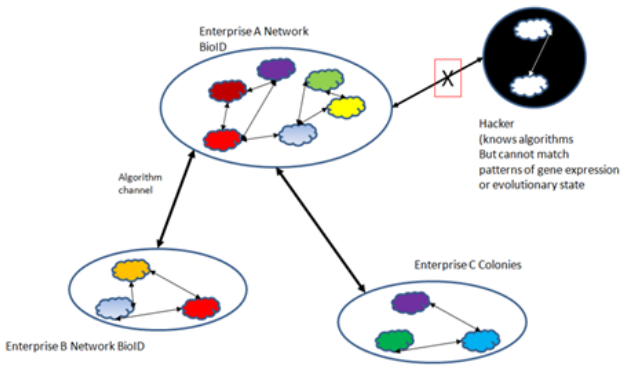

Fig. 4. Network Concept of Operations using regulation of gene expression

Networks would have appliance, called a Network BioID. Enterprises will maintain Network BioIDs that communicate with each other via communication channels (the algorithm channel). The ciphercolonies are regularly sampled to measure their patterns of gene expression and receive external stimuli to modify those patterns of expression. Patterns of gene expression occur with the expression of multiple genes which interact with each other. Gene A, produces protein A, which effects the expression of gene B, which produces protein B, etc. This creates a series of gene regulatory networks. *In vivo*, these regulatory networks appear in close proximity to each other, within the same cell, or colony of cells through a process of cellular signaling. In a computer network, that proximity need not be a physical proximity, but a communications channel. *In vivo*, the genes all exist in a physical sense within the nucleus or cellular compartment housing the DNA. In a computer network, some genes can exist in a virtual sense and these genes can communicate with the physical ones via a signaling process which alters the patterns of expression in both algorithmic and live participants.

Security rests upon maintaining continual knowledge of the state of gene expression of the underlying colonies and their interactions. For an initial implementation, it is possible to monitor and record patterns of gene expression in a laboratory over a long baseline period of time, (e.g. 1 year) and use that recorded data in place of real-time knowledge of patterns of gene expression. Additional data would be collected in parallel with consuming the recorded data is being consumed, while maintaining a one year reserve of gene expression data.

C. *Incorporation into legacy networks*

A new concept such as the ones proposed in this research will not be accepted unless it can be interfaced into legacy, non-bio capable networks. In fig.5. Alice and Bob are shown using genomic and proteomic authentication in conjunction with IPsec. In this case Alice and Bob have Network BioIDs incorporating algorithmic and live sources of gene expression and a verification method using fluorescence pattern matching when gene expression is forced by appropriate stimuli.

Alice and Bob can perform IPsec using nonces and keys derived from patterns of gene expression in their ciphercolonies unknown to them and to Eve. Alice and Bob can perform authentication via messages directing the receiver to force expression of gene(s) which can be detected optically via fluorescence, successfully as in case (a), or unsuccessfully as in case (b). Alice and Bob can send genomically or proteomically encrypted messages that cannot be decrypted without knowledge of the patterns of gene expression used as the basis of encryption. Every network node can be equipped with a Network BioID. Eve will need a background in molecular biology and cryptography as well as lab resources to attack this protocol. Coded patterns of fluorescence detection of gene expression can be formed and utilized with current technology. This type of security will require organizations to take on functions and capabilities very different from those currently used in IT departments.

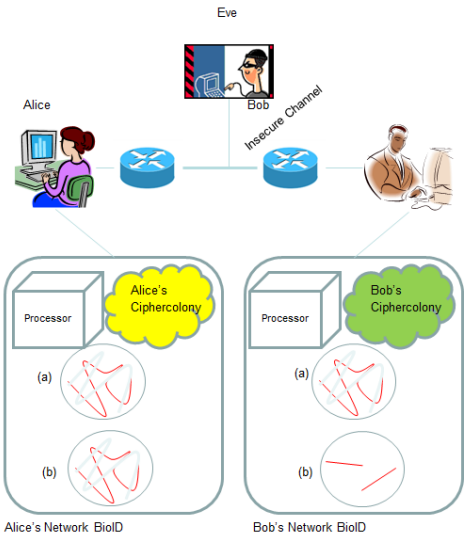

Fig. 5. Combined legacy and genomic protocol security

D. *Proteomic Authentication Messages*

An IT security official receives a remote request for access to network assets from a remote user. The security official sends the user a message coded as a protein sequence, by a regulatory network using a message-specific set of protein-DNA Type codes and a source coding scheme based upon a keyed hash function tied to a specific genome. The user successfully decrypts the message and returns the plaintext (which could be encrypted if desired) to the IT security official. The IT security official then sends a set of access credentials encrypted with a different protein and a different genome for the keyed hash code. The user successfully decrypts the message to gain access to the network. In this scheme, an attacker needs multiple levels of information at the genomic and proteomic levels to be able to decode the message by cryptanalysis means alone. The process is summarized in fig. 6.

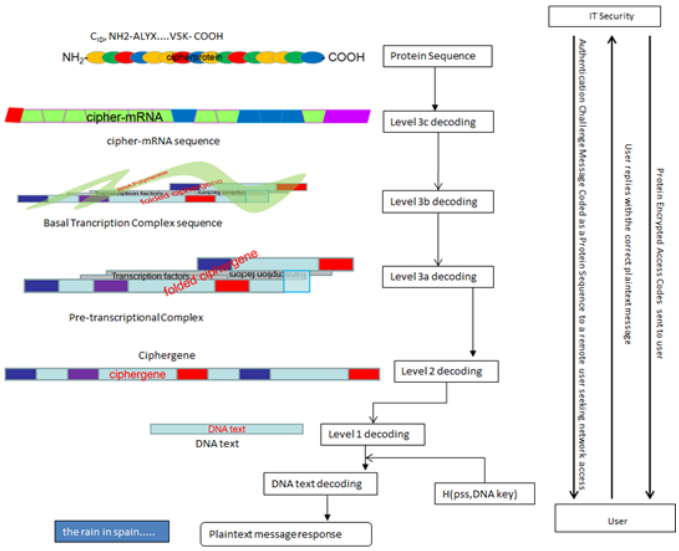

Fig. 6. Protein Coded Authentication Challenge

IV. Aspects of Practical Implementation

A. *Technologies for integrating the protocols with laboratory analysis of gene expression*

The US Patent that has been secured on the technology is forward looking and assumes that full implementation will involve future technological advancements that will enhance the implementation concepts in the form of laboratory-on-a-chip devices [4]. The goal of real-time, lab-on-a-chip assays for gene expression has already been realized [5]. It is also currently possible to use techniques such as dielectrophoretic separation of assay products to create signatures of gene expression which can be used to create cryptographic codes with the genomic and proteomic protocols and low-cost devices such as the one shown in fig. 7 [6].

Fig. 8 is a concept of operations for implementing the protocols within a legacy-style PKI system. The purple boxes on each slide refer to blocks in the BioID ciphercolony which may or may not be local to the user computer performing encryption. The ciphergene ID (CID) is essentially an index that points to the name of gene whose sequence, transcription, and translation features are used in the encryption process. The DNA text message is embedded into the gene sequence by the source coding protocol previously described. A given message can be encoded differently by inserting it into different genes.

B. *Example of implementing protocols into a Public Key Infrastructure using a Bio-Certificate Authority*

The Sender decrypts the GSK with its private key and retrieves the locus control region key (Bio-LCR) from the BioID ciphercolony database. The Bio-LCR is decrypted with the GSK. The DNA text is encrypted with the Bio-LCR, converting the DNA text to a ciphergene. The CID is encrypted with the public key of the sender and concatenated with the ciphergene. This completes Level 1 encryption. The ciphertext message can be sent to the receiver or sent to level 2 for further processing.

The level 1 process is as follows: The Sender encrypts the CID with a Bio-CA generated public key and transmits the encrypted CID to a remote Bio-CA. The Bio-CA decrypts the CID with its private key and retrieves a Gene Sequence Key Encryption Key (GSK) for the message associated with the CID. The Bio-CA encrypts the GSK with the Sender's public key and transmits the GSK to the Sender.

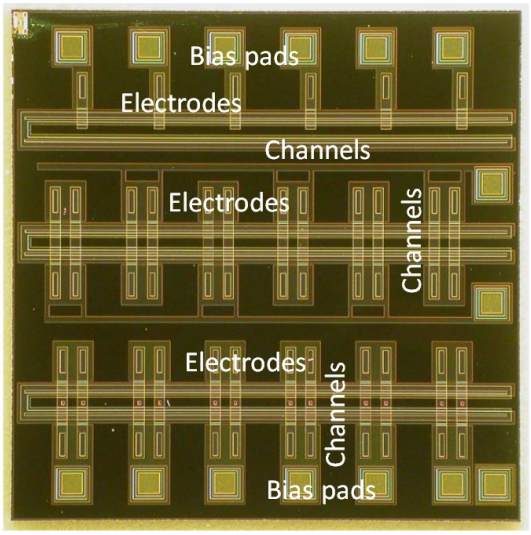

Fig. 7. Dielectrophoretic Separation Test MEMS

REFERENCES

[1] T. M. Cover and J. A. Thomas, , Elements of Information Theory 2nd Ed., pp,103-347, 2006, Wiley Interscience, Hoboken, NJ

[2] S.F. Gilbert, Developmental Biology. 6th edition, Differential Gene Transcription, Sunderland (MA), Sinauer Associates, 2000.

[3] Genomics and Proteomics Based Security Protocols for Secure Network Architectures, Shaw, Harry Cornel, Doctoral Dissertation, The George Washington University, 2013

[4] H. Shaw, "Integrated Genomic And Proteomic Security Protocol," U.S. Patent: 8,898,479, issued date November 25, 2014.

[5] King, Kevin R. and Wang, Sihong and Irimia, Daniel and Jayaraman, Arul and Toner, Mehmet and Yarmush, Martin L., "A high-throughput microfluidic real-time gene expression living cell array", The Royal Society of Chemistry,Lab Chip, Vol. 7, Issue 1, pp. 77-85, 2007

[6] H. C. Shaw, "Design And Simulation of a Mems Structure for Electrophoretic and Dielectrophoretic Separation of Particles by Contactless Electrodes," M.S. thesis, Electrical and Computer Engineering, George Washington Univ., Washington, DC, 2005

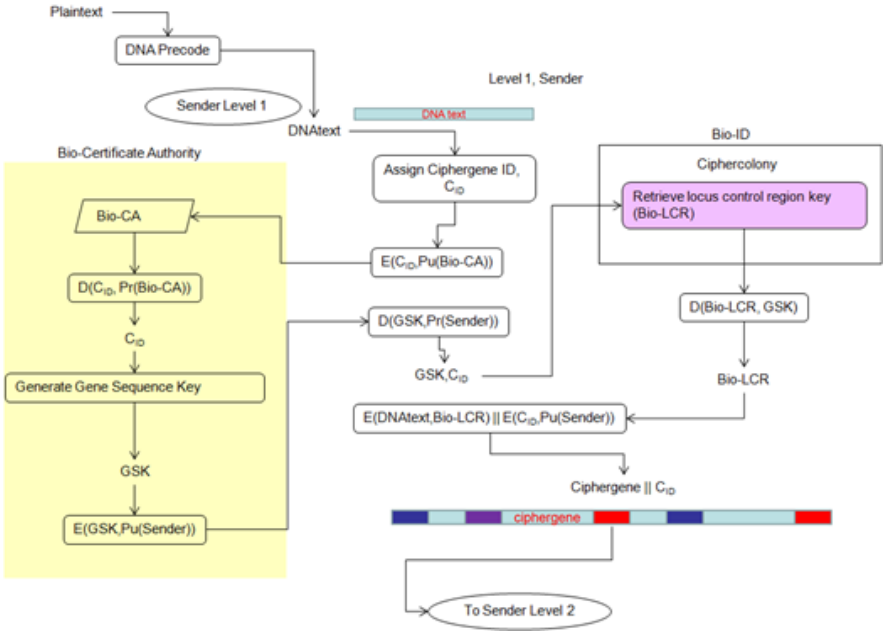

Fig. 8. A Public Key Infrastructure implementation using Bio-Certificate Authorities