National Aeronautics and Space Administration

NASA

ESC Exploration & SPACE Communications

More than you ever imagined...

Genomics and Proteomics Based Security Protocols for Secure Network Architectures

Dr. Harry Shaw
NASA/Goddard Space Flight Center
U.S. Patent 8,898,479, 2014

What is it?

- A hardware design that integrates live and algorithmic inhabitants to produce patterns of gene expression *in vivo* and *in silico*
- Protocols and algorithms based upon the processes of regulation of gene expression to produce cryptographic representations of genes, RNA, proteins, and gene expression to perform authentication and confidentiality functions for computers and networks
- A network concept of operations integrating all of the above into existing legacy networks

Business Applications for the Integrated Genomic Proteomic Security Protocols

- Subscription service for smartphone security app
  - Peer-to-peer would be the simplest implementation,
  - Personal authentication services (digital bio-signatures)
- SaaS Subscription services
  - Data at Rest, Data in Motion, Critical Network restoral data, Data integrity services
- Corporate Data Centers/Cloud data security
- Trusted 3rd party (Bio-Certificate Authorities)
- Digital Rights Management
- Network-Network Authentication, Virtual Private Networks
- Bio-firewalls

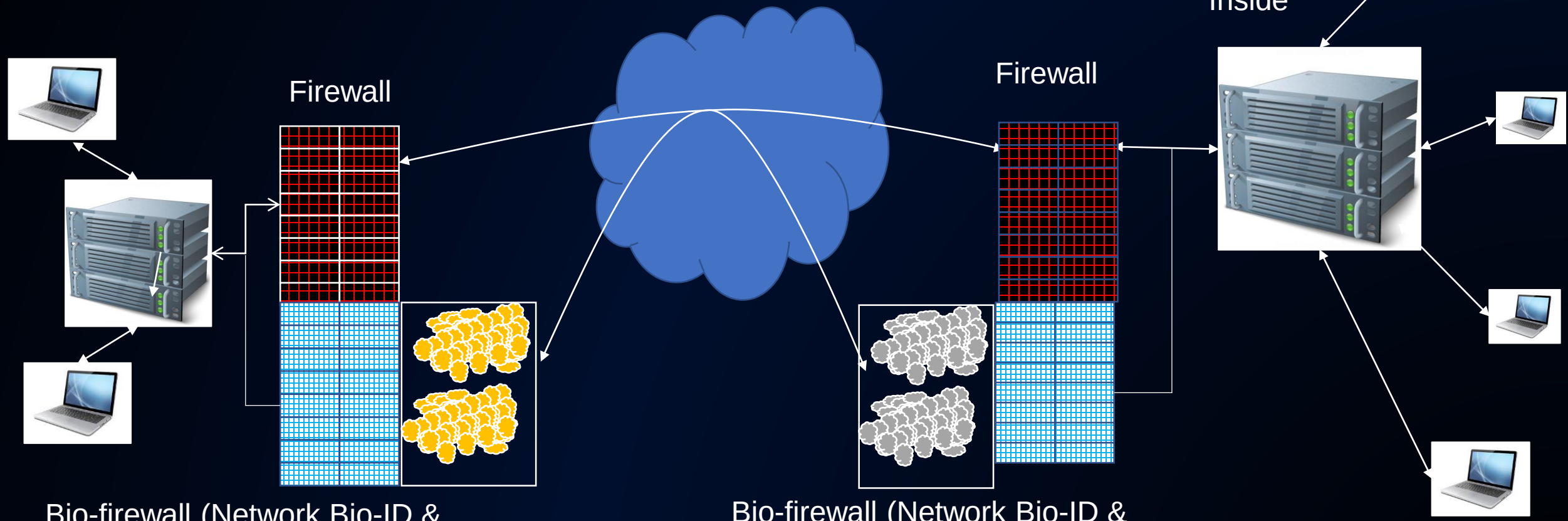
Packet filtering

Connection state of packets
Application firewalls
Accept, reject, or drop based
Source or destination

Outside

Inside

Firewall



Bio-firewall (Network Bio-ID & Ciphercolony)

Do I recognize you?
Did I know you in the past?
What have we shared?

Packet filtering

Connection state of packets
Application firewalls
Accept, reject, or drop based
Source or destination

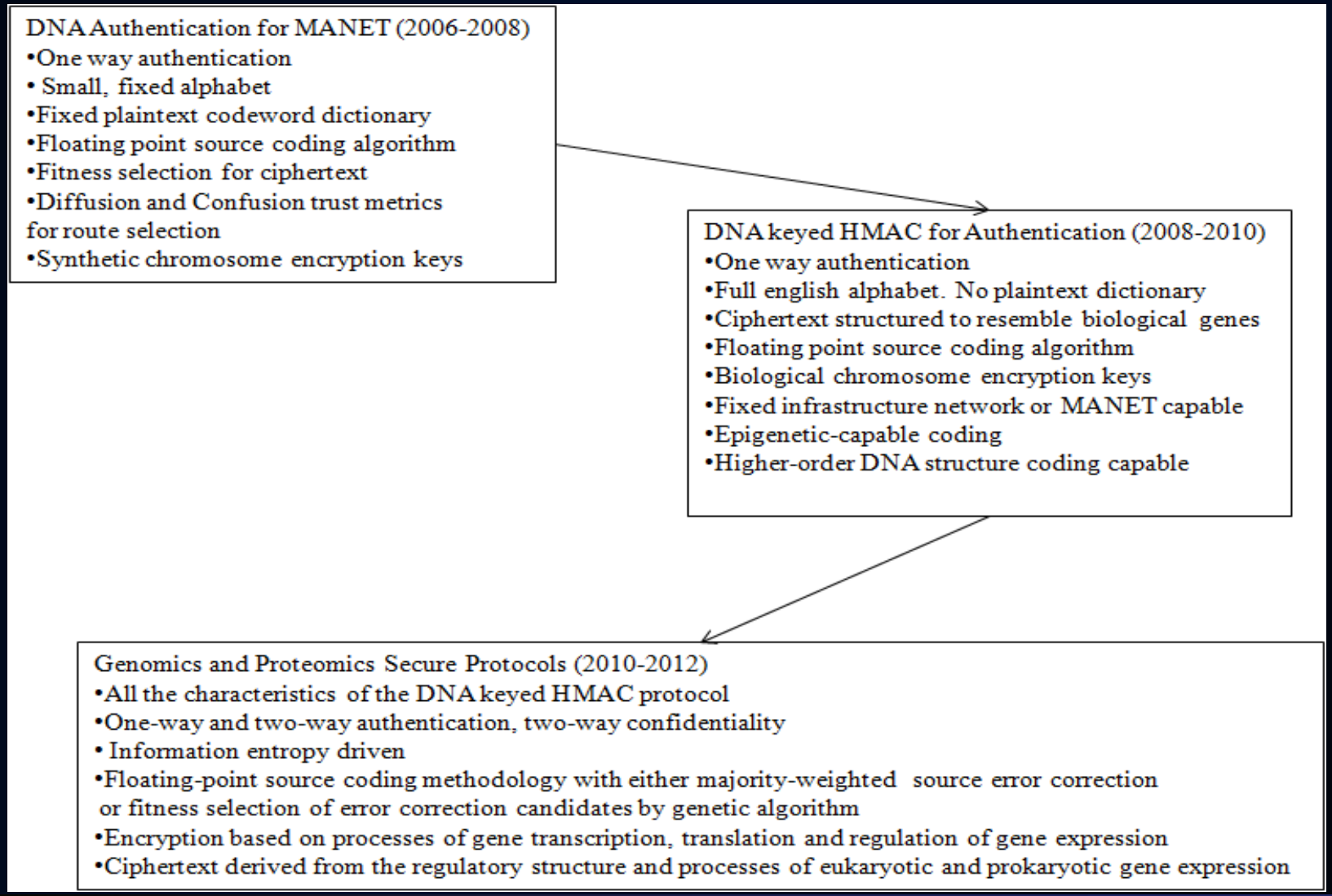
Inside

Firewall

Bio-firewall (Network Bio-ID & Ciphercolony)

Do I recognize you?
Did I know you in the past?
What have we shared?

Research Progression



Motivation for this Research

- 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.
- Credible alternative concepts face a high barrier to entry in the IT security market due to the perceived cost and effort of implementation

Definition of Insanity - Repeating the same action over and over and expecting a different outcome

Why Genomics and Proteomics?

- Genomics and proteomics involve modelable networks which can be converted into cryptographic codes at many levels.
  1. Nucleic acid – protein level (networks of nucleic acid-protein interactions and nucleic acid – nucleic acid interactions for regulation of gene transcription and translation)
  2. Patterns of gene expression (networks of gene interactions)
  3. Intercellular systems (networks of cellular interactions, e.g. biofilms)
  4. And so forth (complex eukaryotic and prokaryotic systems)

Most of today's discussion is at Level 1

Background: What are the “omics” and the “omes”

- There is a plethora of “omics” in the lexicon. In this talk, I will refer to genomics, proteomics, transcriptomics
- Genomics is the study of the set of genes in genome, their functions, interrelationships, and characteristics
- Proteomics, is the study of the set of proteins derived from a genome, (the proteome) their functions, interrelationships and characteristics
- Transcriptomics is the study of the transcriptome, the products of transcription which are RNA their functions, interrelationships, and characteristics

Fundamental premise of the protocols

1. Every plaintext message can be converted to a representation in DNA (why DNA?)
2. The DNA text is operated on by codes representing the interactions of:
   a. Proteins on proteins
   b. Nucleic acids on proteins
   c. Nucleic acids on nucleic acids
3. Encryption is based upon the processes of gene expression using transcription and translation using the interactions described above in (2.)
4. The set of all interactions represents a cryptographically hard scheme which can be used for
   a. Confidentiality
   b. Authentication
   c. Data Integrity
   d. Access Control
   e. Non-repudiation

Relationship between Cryptography and the “omics”

- The ability to authenticate the identity of participants in a network is critical to network security. Bimolecular systems of gene expression “authenticate” themselves through various means such as transcription factors and promoter sequences.
- They have means of retaining “confidentiality” of the meaning of genome sequences through processes such as control of protein expression.
- These actions occur independently of a centralized control mechanism.
- The overall goal of the research is to develop practical systems of authentication and confidentiality such that independence of authentication and confidentiality can occur without a centralized third party system.

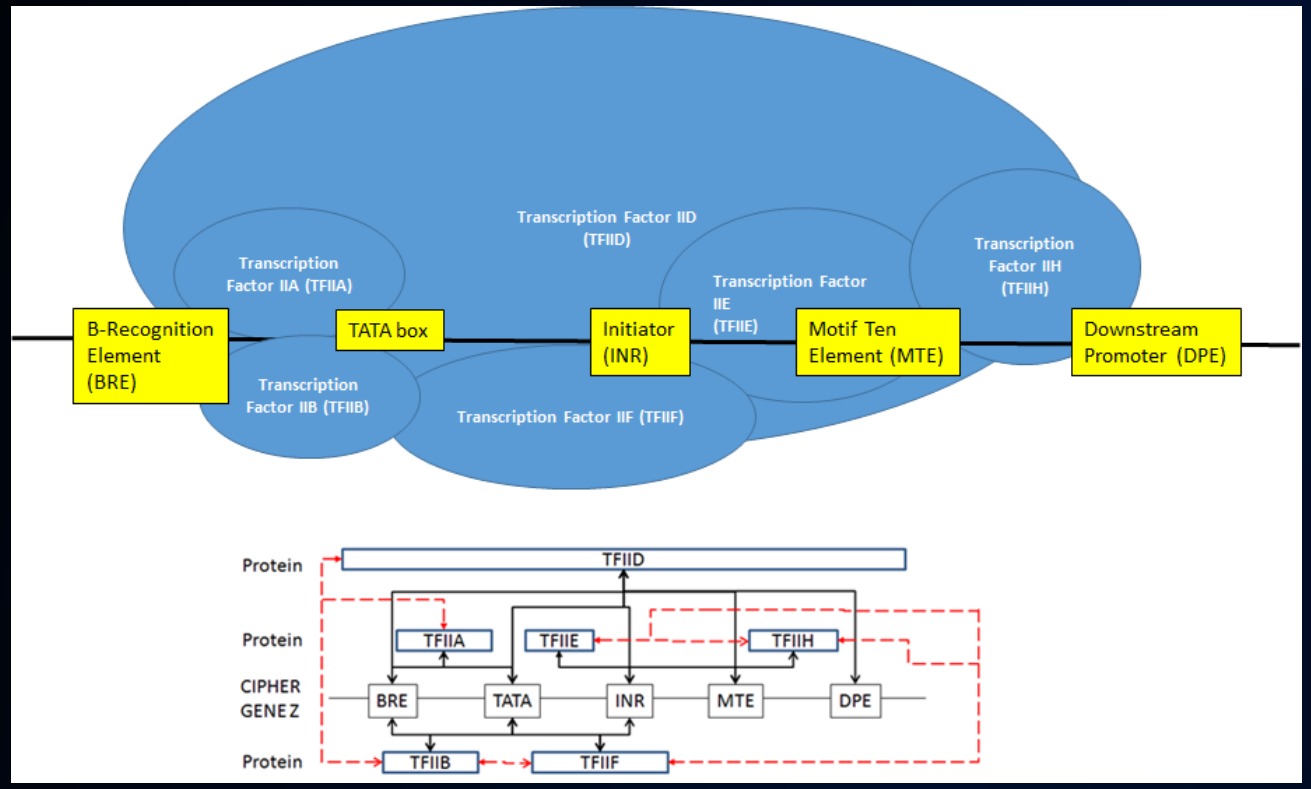
Overview of the Cryptographic protocol

Encryption	Level	Input	Output	Decryption	Level	Input	Output
↓↓↓	-	Plaintext	DNA text	↓↓↓	3C	Cipherprotein	c-mRNA
↓↓↓	1	DNA text	Ciphergene	↓↓↓	3B	c-mRNA	BTC
↓↓↓	2	Ciphergene	Pre-transcriptional complex (PTC)	↓↓↓	3A	BTC	PTC
↓↓↓	3A	PTC	Basal transcriptional complex (BTC)	↓↓↓	2	PTC	Ciphergene
↓↓↓	3B	BTC	Cipher messenger RNA (c-mRNA)	↓↓↓	1	Ciphergene	DNA text
End	3C	c-mRNA	Cipherprotein	End	-	DNA text	Plaintext

Organization of genes

- Genes contain regulatory sequences that control expression through binding of proteins instead of expression of products through transcription
- Binding to these sequences can enhance or reduce levels of expression
- The types of sequences include types such as
  - Promoters
  - Enhancers
  - Silencers
- They are generally short sequences (up to 200 base pairs) but recognition of these sequences is required for proper levels of gene expression

General Transcriptional Complex and its representation

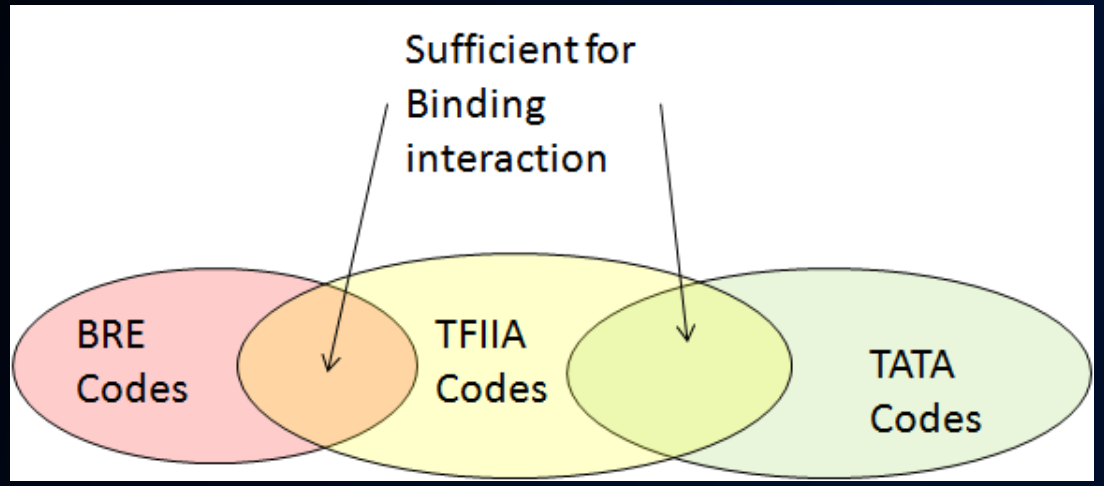


The square blocks are regulatory sequences in a gene.
These sequences are required to bind the transcription factor Proteins required for transcription
The ovals are general transcription factor proteins. They are Required to bind to the regulatory sequences so the RNA Polymerase II can effect transcription of the gene into RNA

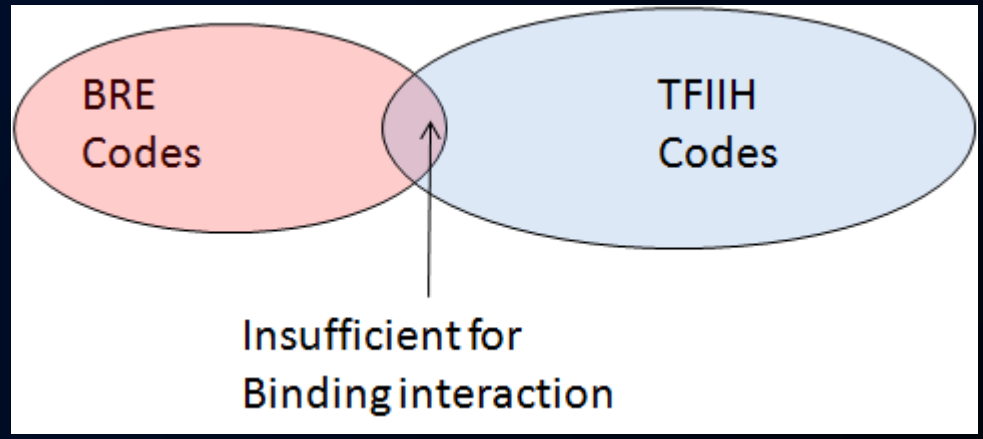
The red dotted lines are protein-protein interactions
The solid black lines are protein-nucleotide interactions

Probabilistic binding conditions

Binding of BRE to TFIIA and TATA to TFIIA



Non-binding of BRE to TFIIH



For example, there exists a condition of binding such that codes from BRE and TFIIA and TATA and TFIIA satisfy a condition at a binding threshold.

$$J = P(BRE \cap TFIIA) \cap P(TATA \cap TFIIA)$$

Coding for Control of Transcription Factor Binding An example

Let $\Gamma = \{1, 2, 3, 4, 5\}$, a 5-tuple alphabet for gene regulatory sequences with type Γ_g consisting of

$$P_{g1} = (2/10) = 0.2$$
$$P_{g2} = 0.4$$
$$P_{g3} = 0.1$$
$$P_{g4} = 0.1$$
$$P_{g5} = 0.2$$

$$T(\Gamma_g) = \{11222345,112225534,\ldots,5543222211\}$$
$$\left|T\left(\Gamma_g\right)\right| = \left(\frac{10!}{2!4!2!}\right) = 37{,}800$$

BRE as $g = 2455222113$ as a member of the type Γ which can contain all the codes for those regulatory sequences

Let $\Psi = \{0,1,2,4,5,8,9\}$ 7-tuple alphabet of transcription factor codes for members of TFIIx (TFIIA, TFIIB, etc.)

Let Ψ_{tf} consist of sets that conform to

$$P_{\Psi 0} = 1/10$$
$$P_{\Psi 1} = 1/10$$
$$P_{\Psi 2} = 2/10$$
$$P_{\Psi 4} = 2/10$$
$$P_{\Psi 5} = 1/10$$
$$P_{\Psi 8} = 1/10$$
$$P_{\Psi 9} = 2/10$$
$$\left|T\left(\Psi_{tf}\right)\right| = \left(\frac{10!}{2!2!2!}\right) = 453{,}600$$

TFIID as $tf = 5089292414$ fits the condition which contains the codes for all of the transcription factor proteins

Joint distribution of gene regulatory sequences and transcription factor codes

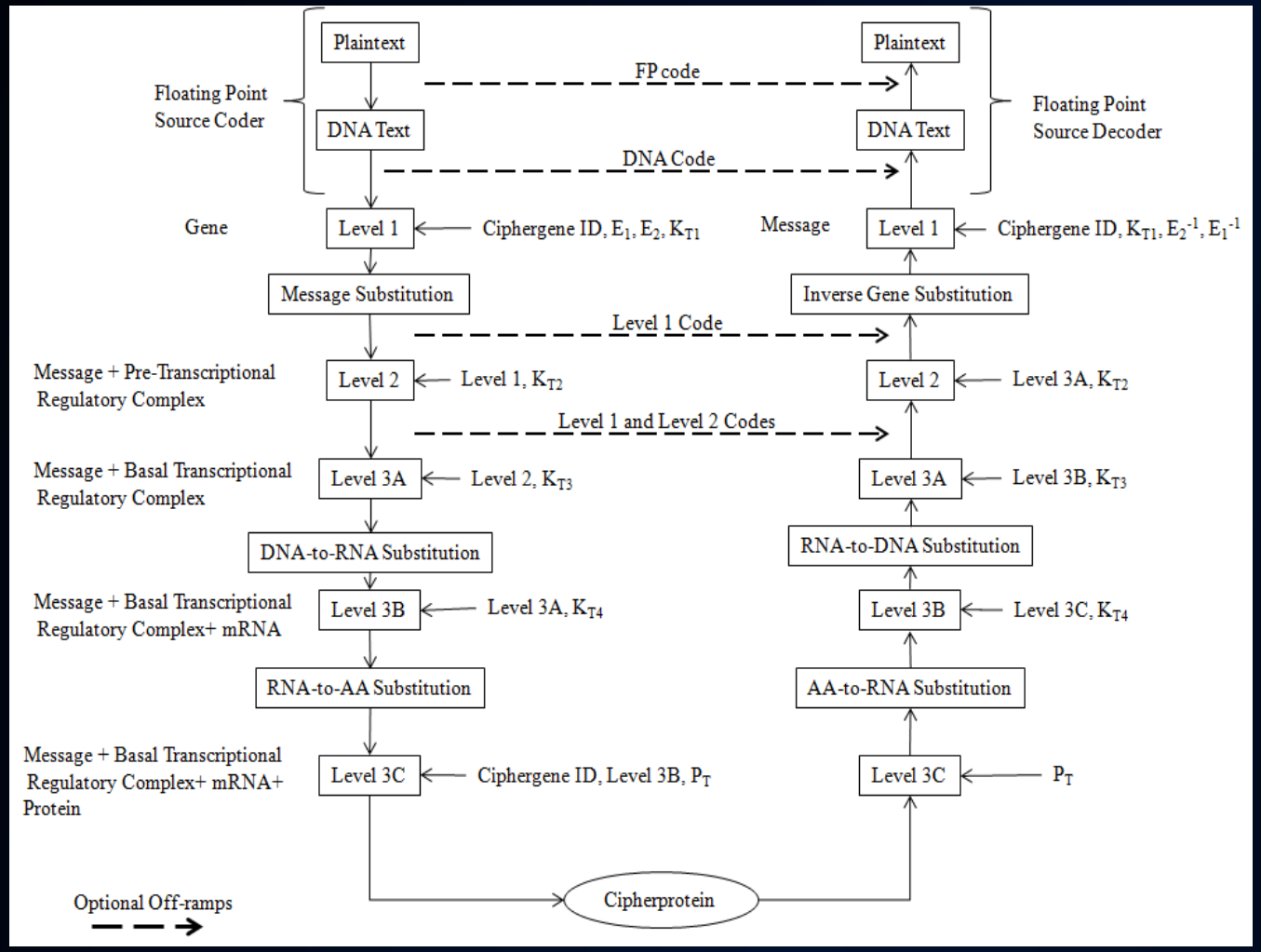
	tf	0	1	2	4	5	8	9
g		0.1	0.1	0.2	0.2	0.1	0.1	0.2
1	0.2	0.02	0.14	0.04	0	0	0	0
2	0.4			0.28	0.08	0.04		
3	0.1	0	0	0.07	0.02	0.01	0	0
4	0.1	0	0	0	0.08	0.01	0.01	0
5	0.2	0	0	0	0	0.14	0.02	0.04

Define a new type, Ω, such that it conforms to the joint distribution of Γ and Ψ as shown above. Using the examples of *BRE* as $g = 2455222113$ and *TFIIA* as $tf = 5089292414$ and the output is a codeword complying with the statistical distribution

The process described is used for all interactions

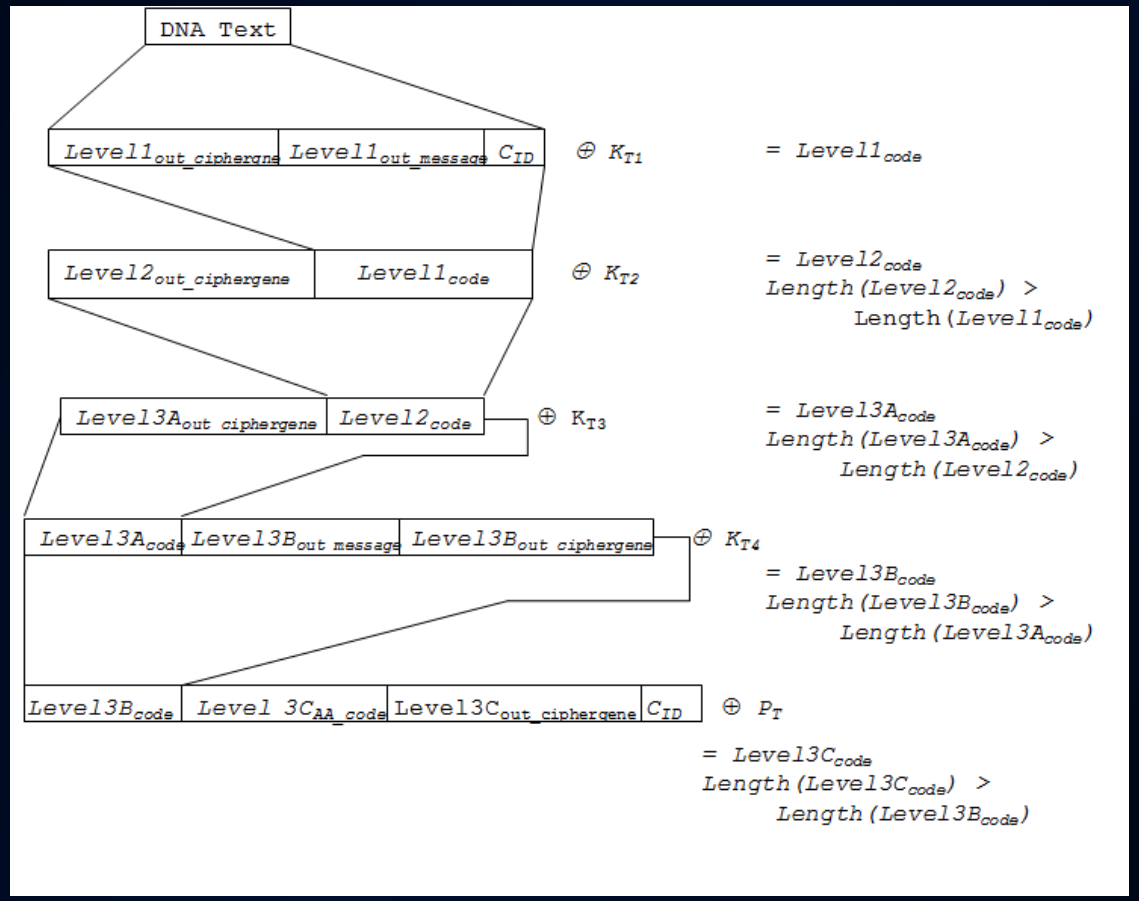
- Define the Types
- Create network diagrams of interactions of the Types
- Define the joint probabilities of the interactions of the Types
- Generate the codewords
- Applies to transcription, translation, regulation of gene expression processes, etc.

Process overview



- Will work with any Plaintext to DNA coding scheme as a starting point
- Optional off-ramps at different encryption levels

Progression of the structure of the ciphertext



Security Features unique to these protocols

Advantages to this approach

- There are millions of processes and combinations of molecular interactions available
- Conventional attack methodologies cannot be applied
- *in vivo* instantiations can result in a new type biological authentication
- Higher levels of abstraction can be coded through networks of gene expression.
- Ciphercolonies can exchange patterns of gene expression as if they were in physical contact as a means of authentication via recognition.

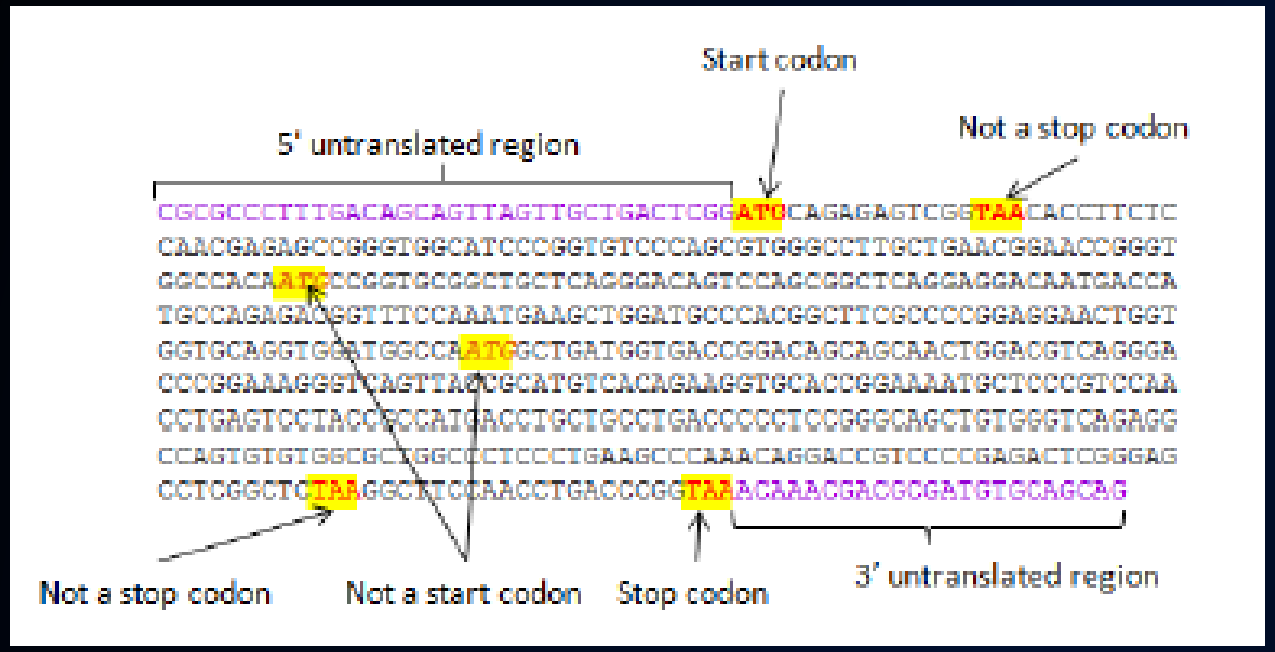
The organization of genes provides for both confusion and diffusion

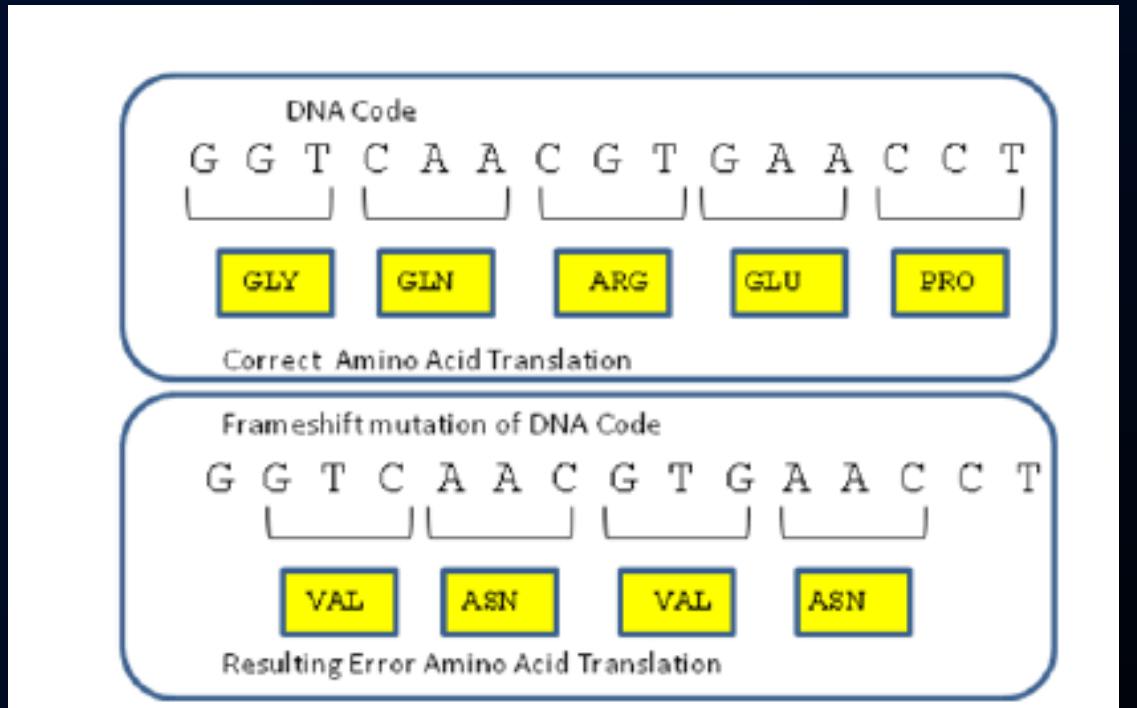
1) Diffusion: any redundancy or patterns in the plaintext message are dissipated into the long range statistics of the ciphertext message.

2) Confusion: make complex the relationship between the plaintext and ciphertext. A simple substitution cipher would provide very little confusion to a code breaker.

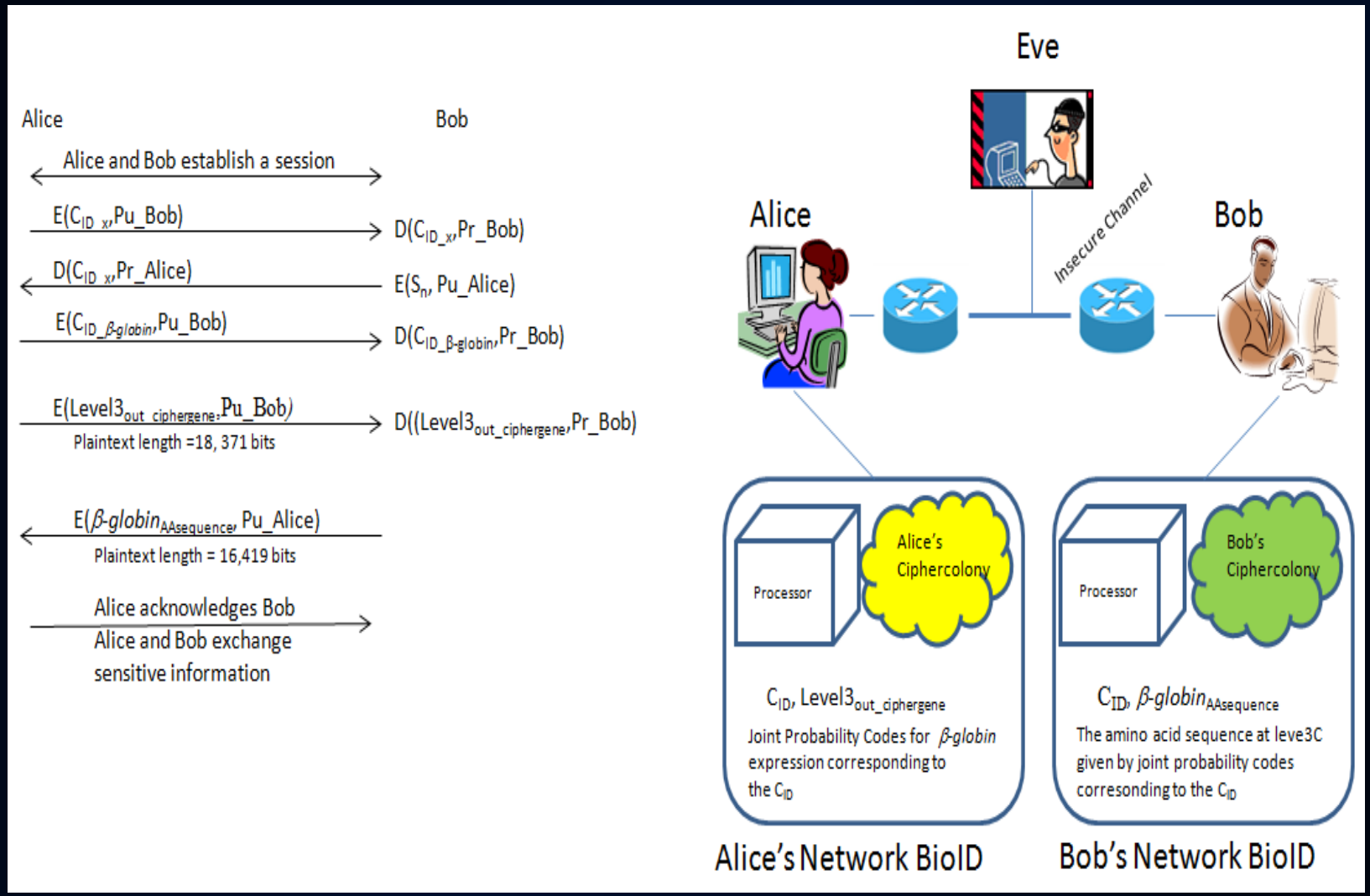
C. Shannon, ―Communication Theory of Secrecy Systems‖,
Bell System Technical Journal, p. 623, July 1948

Elementary confusion and diffusion factors in the organization of the eukaryotic genome





A lightweight implementation example

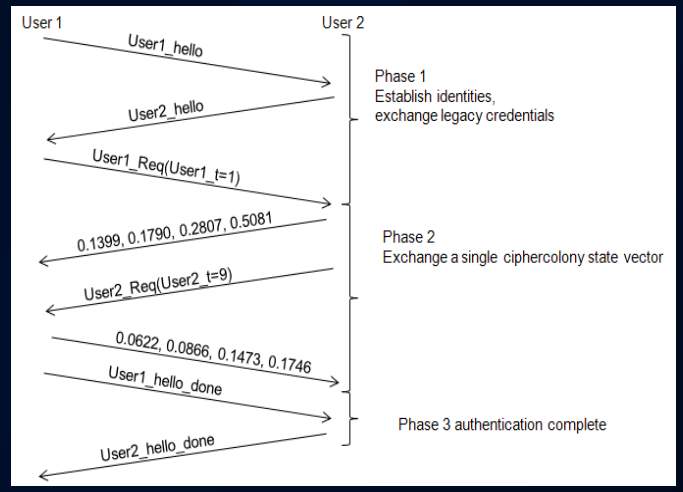


Use of gene expression data for handshaking

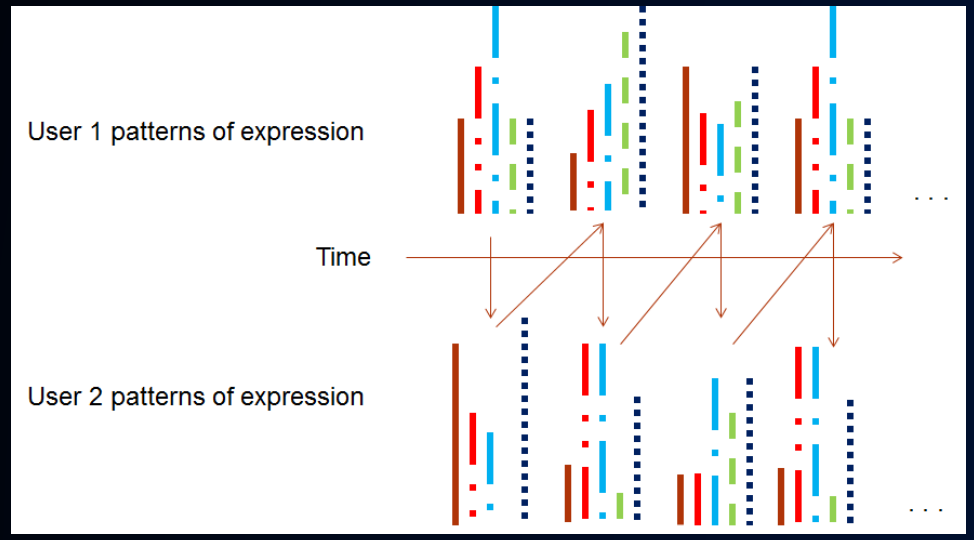
time	User 1				User 2				time
t= 1	0.1399	0.1790	0.2807	0.5081	0.2822	0.2648	0.1694	0.5898	1=t
t= 2	0.5875	0.5675	0.3227	0.8917	0.4332	0.3784	0.2274	0.8706	2=t
t= 3	0.3026	0.3325	0.3154	1.5680	0.3642	0.3208	0.2816	0.8511	3=t
t= 4	0.1446	0.1411	0.1407	0.4850	0.0000	0.0000	0.0000	0.0000	4=t
t= 5	0.3156	0.3009	0.3122	0.5991	0.0000	0.0000	0.0000	0.0000	5=t
t= 6	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	6=t
t= 7	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	7=t
t= 8	0.1015	0.1084	0.1359	0.9908	0.0000	0.0000	0.0000	0.0000	8=t
t= 9	0.2995	0.2711	0.2298	0.6878	0.0622	0.0866	0.1473	0.1746	9=t
t= 10	0.1894	0.1994	0.2551	0.3658	0.2591	0.2742	0.2837	1.6696	10=t
t= 11	0.5196	0.4903	0.2624	0.8183	0.2455	0.2446	0.2685	0.5305	11=t

User 2 State of expression for Proteins A and B

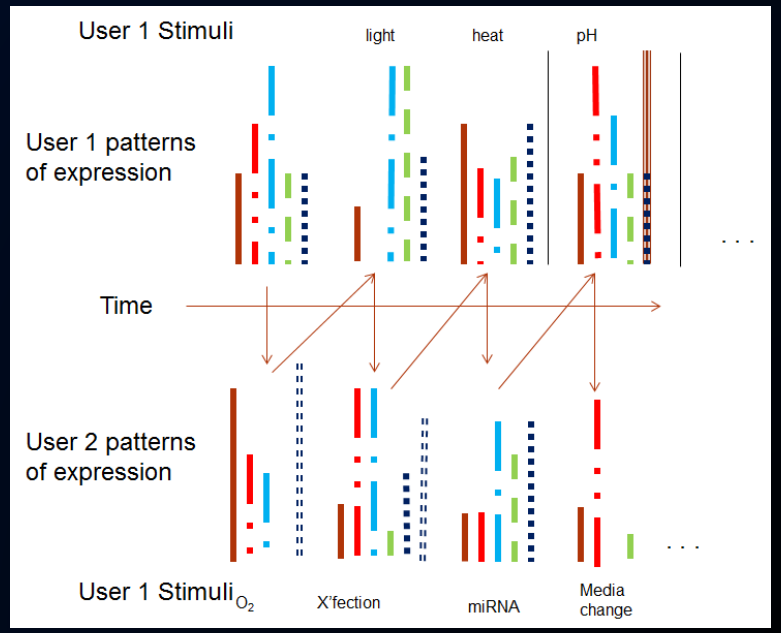
Exchange of ciphercolony state information about Proteins A and B between User 1 and User 2.



Handshaking protocol between User 1 and User 2.

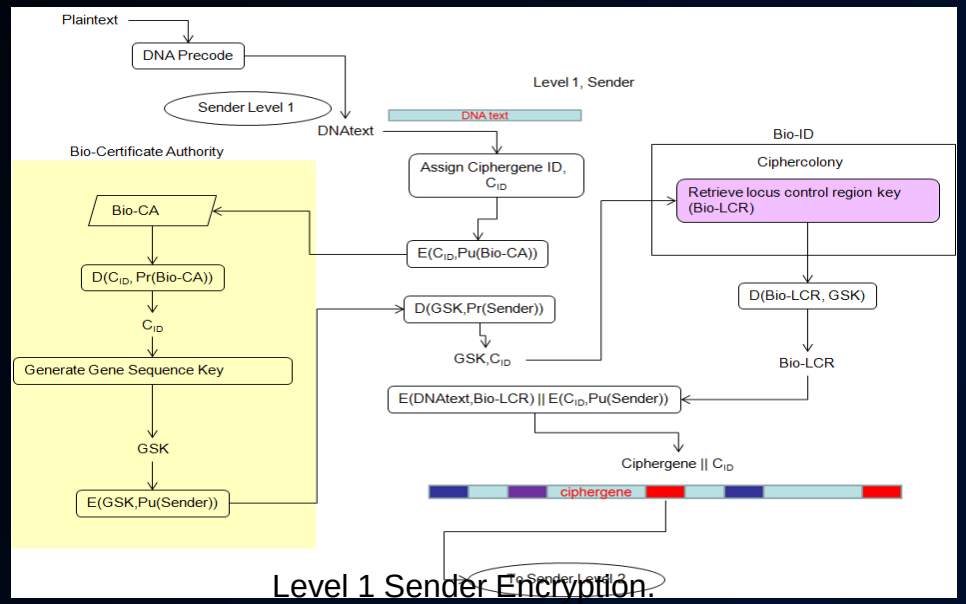


Non-secure

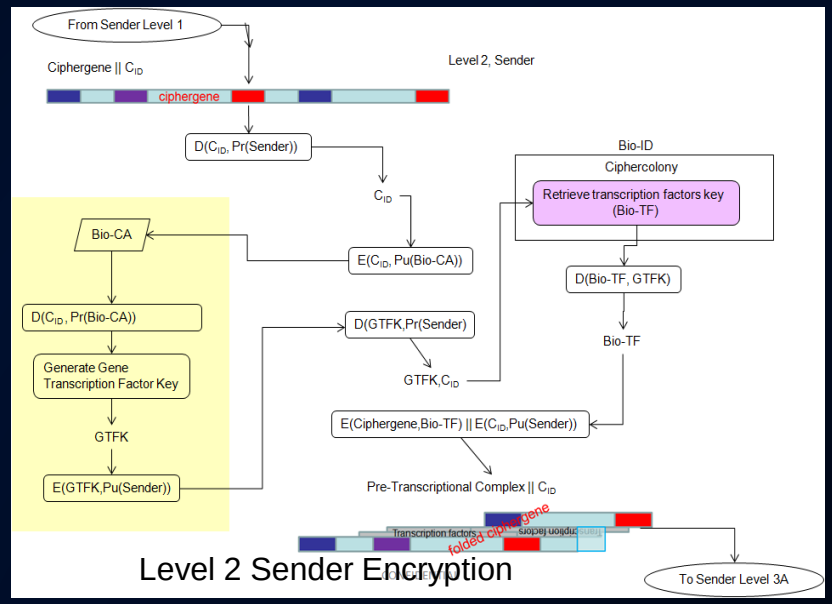


Secure

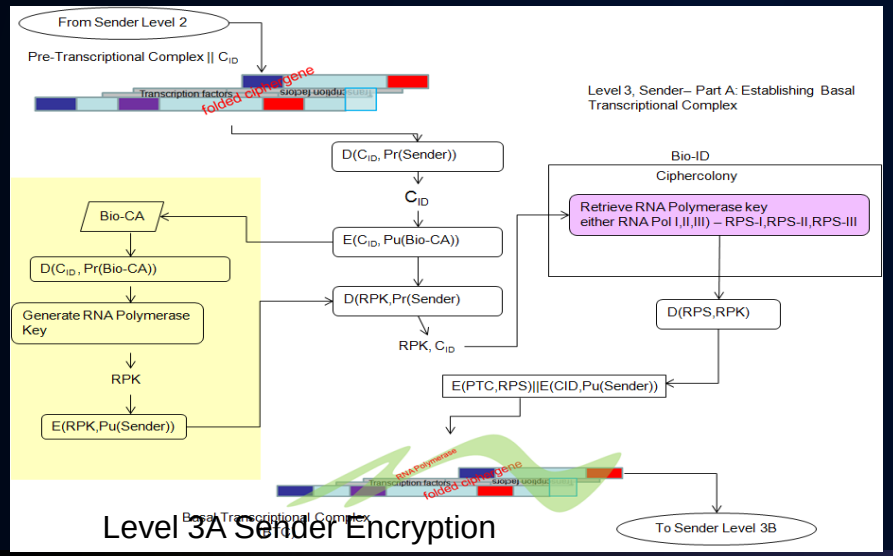
Implementation Options using a Certificate Authority



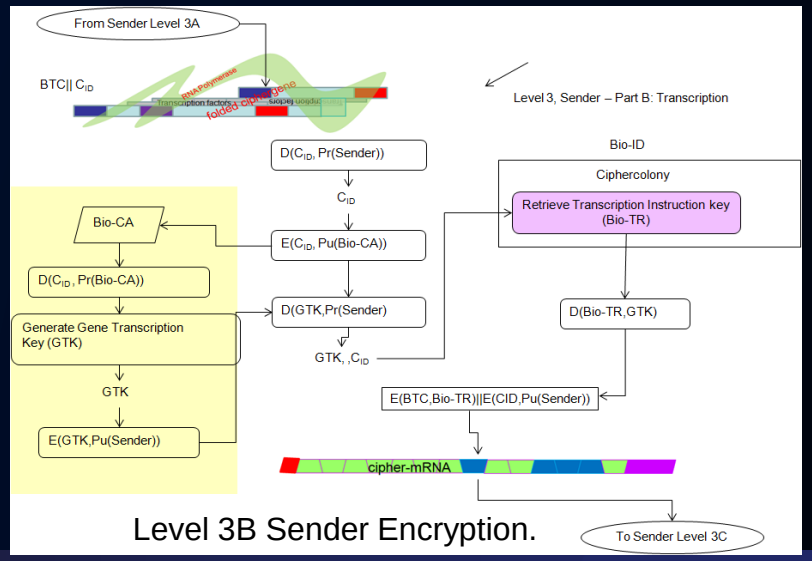
Level 1 Sender Encryption.



Level 2 Sender Encryption

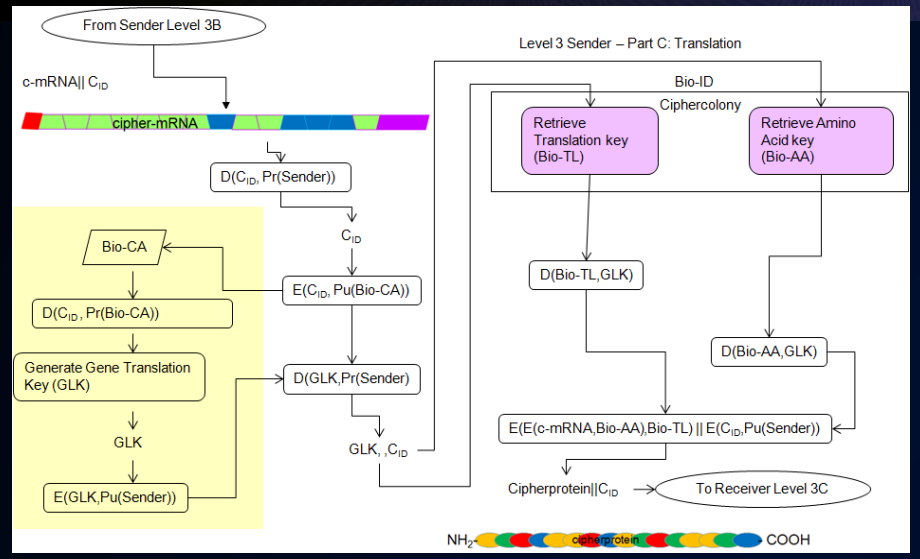


Level 3A Sender Encryption

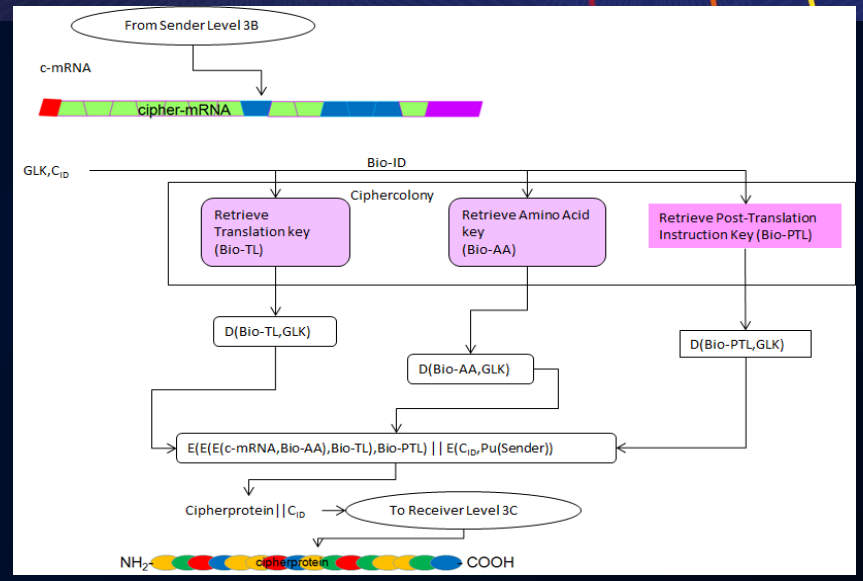


Level 3B Sender Encryption.

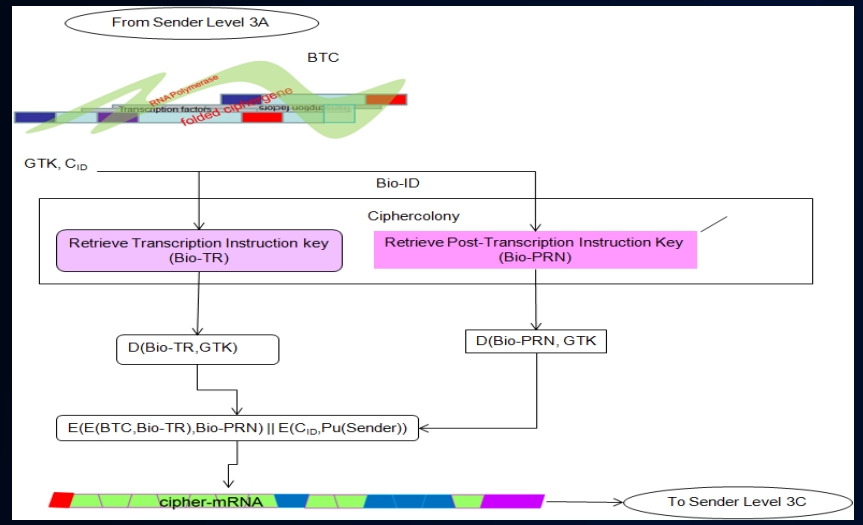
Post-transcriptional, Post-translational and Protein Detection Options



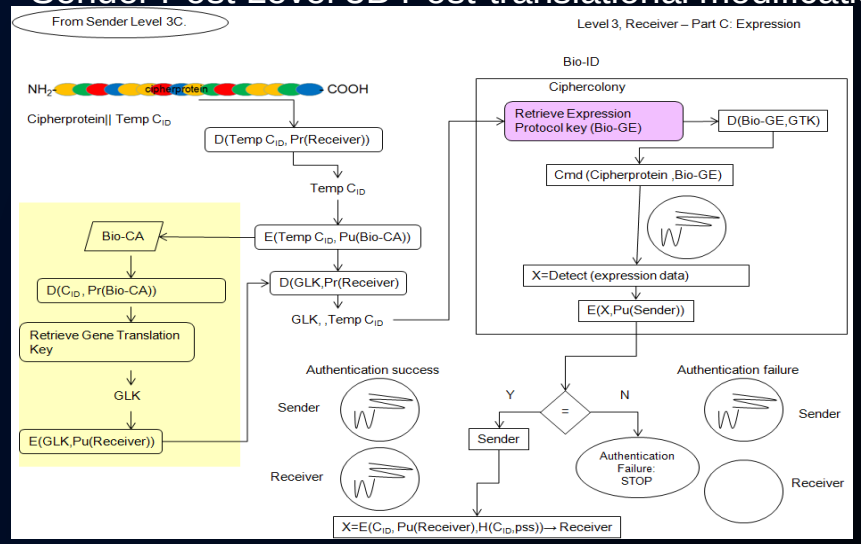
Level 3C Encryption



Sender Post Level 3B Post-translational modifications

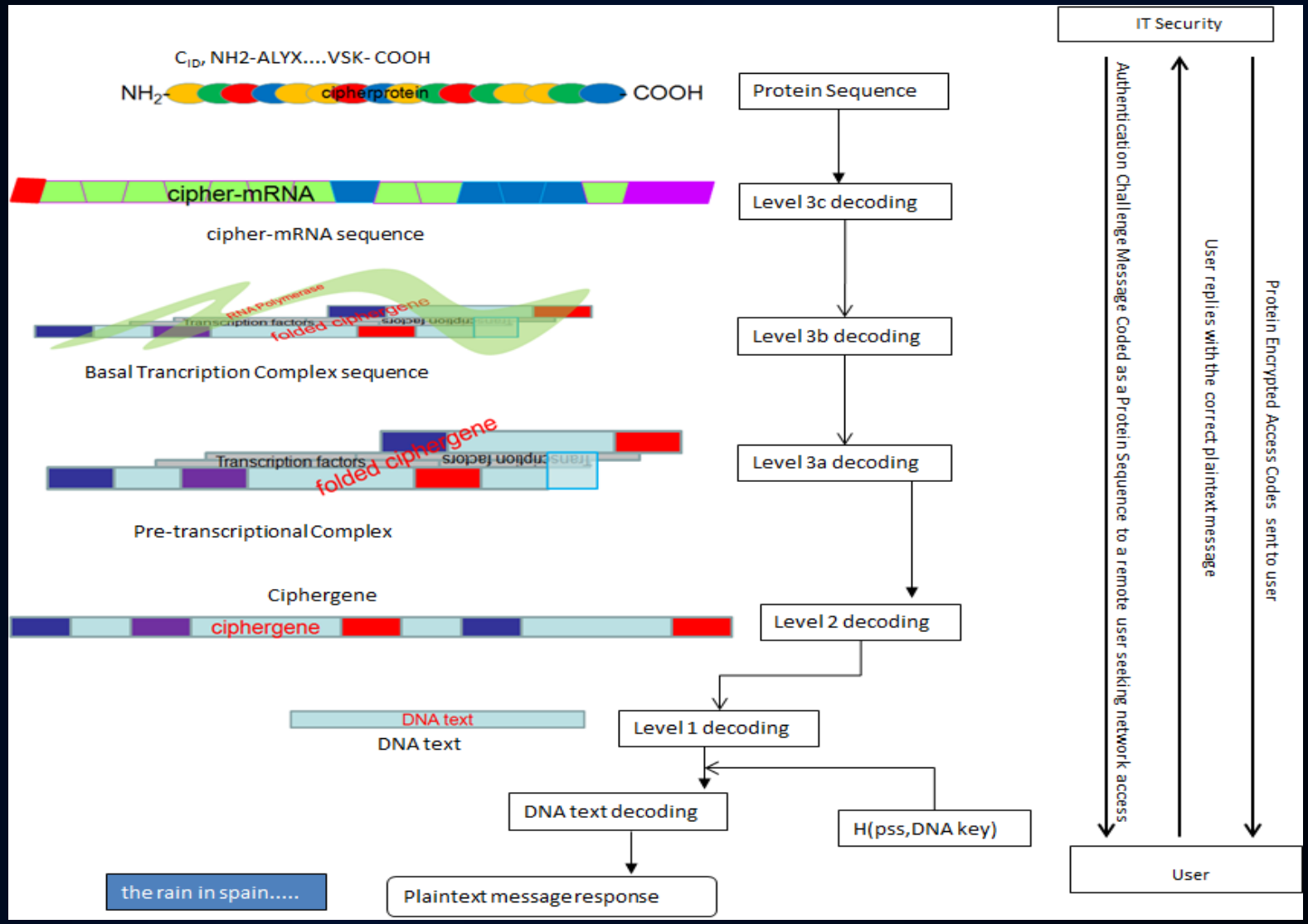


Sender Encryption Post Level 3A Post-transcriptional modifications



Receiver processing with fluorescent protein detection

An authentication challenge using protein codes



Acronyms

- RNA – Ribonucleic Acid
- DNA – Deoxyribonucleic Acid
- HMAC – Hashed Message Authentication Code
- MANET – Mobile Ad Hoc Network
- IT – Information Technology
- National Health Service
- Office of Personnel Management
- PTC - Pretranscriptional complex
- BTC – Basal Transcriptional complex
- mRNA – messenger RNA
- C-mRNA - cipher messenger RNA

Acronyms

- TFIIA – TFIIH: Transciptional Factor IIA through Transcriptional Factor IIH
- BRE: B Recognition Element
- TATA: Thymine Adenine Thymine Adenine
- INR: Initiator
- MTE: Motif Ten Element
- DPE: Downstream Promoter Element
- FP: Floating Point
- AA: Amino Acid