2014년 10월 전자공학회 논문지 제 51권 제 10 호
Journal of The Institute of Electronics and Information Engineers Vol. 51, NO. 10, October 2014 http://dx.doi.org/10.5573/ieie.2014.51.10.118

논문 2014-51-10-16

DNA 데이터 저장을 위한 DNA 정보 은닉 기법

(DNA Information Hiding Method for DNA Data Storage)

이 석 환*, 권 기 룡**

(Suk-Hwan Lee and Ki-Ryong Kwon©)

요 약

DNA 데이터 저장(Data storage)은 DNA의 염기 서열에 대용량의 디지털 데이터를 저장하는 방법으로, 차세대 정보 저장 매개물로 인식되고 있다. 본 논문에서는 DNA 스테가노그라픽 기반으로 비부호 DNA 서열(Noncoding DNA sequence)에 정보를 저장하는 방법을 제안한다. 제안한 방법은 암호화된 데이터들을 정수 변화표에 의하여 데이터 염기 서열로 변환한 후, 시드 정보, 및 섹터 길이로 구성된 은닉 키에 의하여 비부호 염기 서열에 은닉한다. 따라서 단백질의 유전 기능이 유지되고, 원 DNA 서열없이 정보가 검출되며, 변이에 의하여 발생되는 오류가 검출된다. 기존 방법과의 비교 실험을 통하여 제안한 방법이 높은 bpn를 가지는 저장 효율을 가지며, 패리티 염기에 의하여 은닉된 정보의 오류 위치를 검출할 수 있음을 확인하였다.

Abstract

DNA data storage refers to any technique for storing massive digital data in base sequence of DNA and has been recognized as the future storage medium recently. This paper presents an information hiding method for DNA data storage that the massive data is hidden in non-coding strand based on DNA steganography. Our method maps the encrypted data to the data base sequence using the numerical mapping table and then hides it in the non-coding strand using the key that consists of the seed and sector length. Therefore, our method can preserve the protein, extract the hidden data without the knowledge of host DNA sequence, and detect the position of mutation error. Experimental results verify that our method has more high data capacity than conventional methods and also detects the positions of mutation errors by the parity bases.

Keywords : DNA data storage, DNA steganography, Noncoding DNA sequence, DNA information hiding

Ⅰ. 서 론

모든 유전체는 유전 부호(genetic code)를 가지는 분자 구조의 DNA(Deoxyribonucleic Acid) 서열을 가지며, DNA 서열은 적절한 환경 하에서 수 백년 동안 정보를 저장하거나 추출할 수 있다[1]. 즉, DNA 서열은 특정 정보에 따라 분자 구조를 변이함으로써 데이터를 저장하는 저장 매개물로 사용될 수 있으며, 또한 비밀 정보를 저장 및 전송하는 전송 매개물로 사용될 수 있다. 이 때 전자를 DNA 데이터 저장이라 하며, 후자를 DNA 스테가노그라픽이라 한다.

DNA는 뉴클레오티드라는 작은 단위로 구성된 두 개의 가닥들이 역평행 사슬을 가지는 이중 나선 구조로 이루어져 있다. 이 때 뉴클레오티드는 아데닌 (Adenine, A), 티아닌 (Thiamine, T), 시토신 (Cytosine, C), 구아닌 (Guanine, G)이라는 4개의 타입을 가진다. DNA 데

* 정회원, 동명대학교 정보보호학과
(Dept. of Information Security, Tongmyong University)
** 정회원, 부경대학교 IT융합응용공학과
(Dept. of IT Convergence and Application Engineering, Pukyong National University)
© Corresponding Author(E-mail: krkwon@pknu.ac.kr)
※ 본 연구는 2011년도 정부(교육과학기술부)의 재원으로 한국연구재단의 기초연구사업 지원을 받아 수행된 것임(NRF-2011-0023118).
접수일자: 2014년05월13일, 수정일자: 2014년09월01일
게재확정: 2014년09월30일

이터 저장의 장점은 반도체 트랜지스터 내에 1과 0의 조합을 사용하는 대신 A, T, C, G의 뉴클레오티드 타입을 사용함으로써 정보를 저장하는 것이다. 예를 들어, n개의 뉴클레오티드들은 $2n$개의 이진 정보 대신 $4n$개의 조합을 생성한다. 이와 같은 DNA의 데이터 저장 능력과 DNA 시퀀싱(Sequencing) 기술의 발달로 인하여 DNA 데이터 저장 및 DNA 스테가노그라픽은 매우 발전적이고, 유망한 바이오 기술로 인식되고 있다. 최근 Church 등은[2] 비트와 염기와의 일대일 대응 단순 부호를 이용하여 53,400 문자의 HTML 문서와 11 JPEG 사진과 JavaScript 코드들로 구성된 정보를 DNA로 부호화한 다음, 이들을 다중 복제하여 DNA의 1 세제곱 밀리미터 내에 5.5 Petabit를 저장하였다.

DNA 서열은 그림 1(a)에서와 같이 엑손(Exon)의 부호 영역과 인트론(Intron)의 비부호 영역으로 구분된다. 부호 영역을 부호 DNA 서열(Coding DNA sequence, CDS)이라 하고, 비부호 영역을 비부호 DNA 서열(Noncoding DNA sequence)라 한다. 엑손은 단백질로 번역되는 유전체의 부호 영역으로 다음 세대에 유전되는 정보를 담고 있으나, 인트론은 단백질로 번역되지 않는 유전체의 비부호 영역으로 정크 DNA라고도 한다. 정보 저장 또는 데이터 은닉 처리 관점에서 살펴보면, 엑손은 외부 정보에 의하여 유전체의 기능이 변경될 수 있으므로, 데이터 은닉시 코돈 중첩 (Codon degeneracy)을 이용하여야 한다. 이 때, 매우 제한적으로 데이터가 은닉되므로, 정보 저장 및 스테가노그라픽에서는 적합하지 않다. 인트론은 자연적 또는 외부적 요인에 의하여 제거 및 변이가 되더라도 생물학적 유전 기능에는 변함이 없는 정크 정보에 해당되므로, 많은 정보를 담을 수 있다. 그러나 제거와 변이가 쉬우므로, 저작권 보호를 위한 워터마킹에는 적합하지 않다. 따라서 그림 1(b)에서와 같이 비부호 DNA은 DNA 데이터 저장 또는 비밀 통신을 위한 DNA 스테가노그라픽[3]-[6]에 적합하며, 부호 DNA은 저작권 보호를 위한 DNA 워터마킹[7~8]에 적합하다.

DNA 데이터 저장과 DNA 스테가노그라픽의 기술들은 기본적으로 DNA 데이터 은닉 기술에 기반을 두고 있으며, 이에 대한 연구가 진행되고 있다. 1999년 Clelland 등[3]은 변이된 DNA의 마이크로도트를 사용함으로써 비밀 메시지를 전달하는 간단한 DNA 스테가노그라픽을 소개하였다. 이후, 디지털 데이터 저장[4], 오류 정정 방법[5], 데이터의 보안성과 용량 증가[6] 등 DNA 정보의 가독성을 향상시키기 위한 많은 연구들이 수행되어왔다. DNA 스테가노그라픽 외에 변이에 강인한 DNA 워터마킹 기법[7~8]이 연구되어져 왔으나, 이는 부호 영역에 워터마크를 삽입하는 것으로 삽입될 워터마

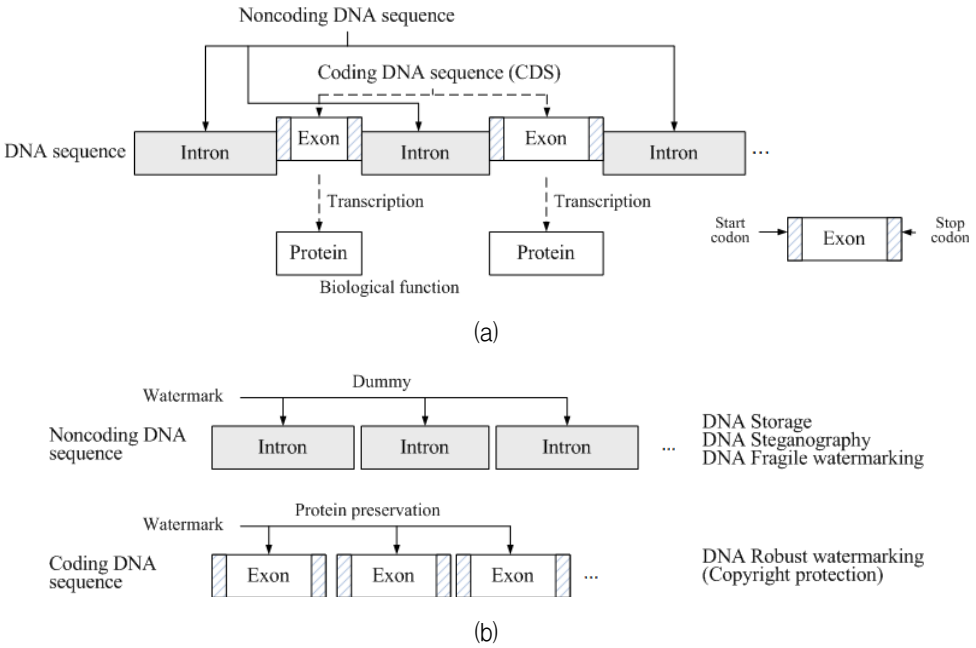

그림 1. (a) DNA 서열 구조 및 (b) 부호 DNA 서열과 비부호 DNA 서열에 대한 워터마킹 응용
Fig. 1. (a) Structure of DNA sequence and (b) applications for coding DNA sequence and noncoding DNA sequence.

크 용량에 매우 제한적이다.

본 논문에서는 DNA의 인트론에 정보를 저장하며, 이 때 보안성, 용량성, 및 오류 검출이 가능한 정보 은닉 방법을 제안한다. 제안한 방법에서는 디지털 데이터를 PRBG(Pseudorandom Bit Generator)에 의하여 암호한 후, 염기 정수 변화표에 의하여 4개의 데이터 염기 서열로 변환한다. 그런 다음, 임의의 길이로 분할된 데이터 염기 서열과 오류 검출을 위한 패러티 정보로 구성된 염기 섹터들을 비부호 영역에 은닉한다. in silico (컴퓨터 내 실험 기술) 실험을 통하여 제안한 방법이 기존 방법에 비하여 약 0.59-1.34 bpn(bit per nucleotide base) 정도 높으며, DNA 서열의 0.1% 변이되더라도 100% 오류 정정됨을 확인하였다. 또한 데이터 은닉된 DNA 서열의 아미노산이 변경되지 않음을 확인하였다.

본 논문의 구성은 다음과 같다. 먼저 Ⅱ장에서는 DNA 데이터 은닉에 대한 기존 연구를 살펴보고, Ⅲ장에서는 제안한 방법에 대하여 자세히 살펴보기로 한다. 그리고 Ⅳ장에서는 제안한 방법과 기존 방법에 대한 실험 결과를 분석하며, 마지막으로 Ⅴ장에서는 본 논문의 결론을 맺는다.

Ⅱ. 관련 연구

최근 DNA 저장 및 스테가노그라픽을 위한 정보 은닉에 대한 연구가 진행되어 왔다. 본 장에서는 대표적인 DNA 정보 은닉에 대하여 살펴보기로 한다.

연구 초창기 Clelland 등[3]은 알파벳 비밀 메시지를 전송하기 위하여 마이크로도트 기술을 사용하였다. 이 방법에서는 CGA='A', CCA='B' 등과 같이 3개의 뉴클레오티드 염기(Nucleotide base)들을 단위로 문자, 숫자, 및 심볼을 할당하는 표를 생성한 다음, 할당 표에 의하여 비밀 메시지를 DNA 서열에 은닉한다. 이 때 PCR 프라이머(Primer)와 비밀 메시지가 은닉되는 시작 마커와 끝 마커는 은닉 키로 사용된다. 그리고 마이크로도트(Microdot)에 숨겨진 은닉 DNA 서열은 은닉 키와 함께 수신자에게 전송된다.

Leier 등[4]은 DNA 서열 상에 은닉되는 이진 정보 부호기를 제안하였다. 이 방법에서는 이진 정보 1 또는 0을 나타내는 짧은 DNA 서열들을 생성한 후, 이들 서열들을 은닉될 이진 정보에 따라 연결한다. 이 때 Clelland 등의 방법과 유사하게 은닉되는 시작 서열과 끝 서열들을 마커하며, 마커 정보들은 은닉 키로 사용된다.

Heider 등[5]은 오류 정정과 높은 삽입 용량을 가지는 DNA-Crypt 알고리즘을 제안하였다. 이 방법은 2비트의 이진 데이터들을 하나의 뉴클레오티드 염기에 "00=T, 01=G, 10=C, 11=A"와 같이 할당한 후, 할당된 염기들을 중첩 코돈 원리에 따라 DNA의 부호 영역에 은닉한다. 이 때 코돈은 3개의 연속된 염기들을 나타내며, 이는 하나의 아미노산으로 번역된다. 아미노산에는 1, 2, 3, 4, 또는 6중 중첩 코돈들 가지는 22개 아미노산들이 있다. 중첩 코돈에 따라 하나의 뉴클레오티드 염기가 중첩 코돈을 가지는 염기로 변경되더라도 아미노산의 변경은 없다. Heider 등은 각 코돈의 마지막 뉴클레오티드 염기를 메시지의 2비트에 따라 변경하며, 이 때 은닉된 코돈 서열들은 원 코돈의 아미노산과 동일한 아미노산을 가지도록 한다. DNA 서열에 대한 변이는 자주 발생되지는 않으나, 은닉된 정보의 오류를 유발한다. Heider 등은 변이에 대한 오류 정정을 위하여 해밍 코드(Hamming code) 또는 반복 오류 정정 부호를 퍼지 조정(Fuzzy control)에 의하여 구현하였다.

Shiu 등[6]은 생물학적 관점보다는 신호처리 관점에서 DNA 서열에 이진 정보를 은닉하는 삽입(Insertion) 방법, 상보형 쌍(Complementary pair) 방법, 대체(Substitution) 방법의 세 가지 방법들을 제안하였다. 첫 번째 삽입 방법에서는 DNA 서열을 변환 규칙에 의하여 이진 배열로 변환한 후 k 비트별로 분할한다. 각 메시지 비트들은 각 DNA 분할 구간 앞에 있도록 부호화된 후 뉴클레오티드 문자들로 다시 변환됨으로써 은닉된 DNA 서열이 생성된다. 상보형 쌍 방법에서는 DNA 사슬에서 가장 긴 서열이 될 유일한 뉴클레오티드 염기 서열을 생성한 후, 이의 상보형 서열을 생성한다. 이 때 메시지 비트는 삽입 방법과 동일하게 뉴클레오티드 문자로 변환되며, 변환된 염기 서열은 상보형 쌍들의 전, 후 염기에 중첩이 되지 않도록 추가된다. 대체 방법은 원 DNA 서열의 랜덤한 위치에 메시지의 이진 정보를 은닉한다. 이 때 메시지 비트가 0일 경우, 해당되는 위치 상의 뉴클레오티드 염기는 변하지 않는다. 이와 반대로 메지시 비트가 1일 경우, 해당 위치에 뉴클레오티드 염기는 상보형 쌍의 염기로 변경된다.

Ⅲ. 제안한 DNA 정보 은닉

제안한 방법은 DNA 스테가노그라픽 기반으로 이진 정보를 DNA 서열에 부호하는 것으로, 우수한 보안성과 용량성 및 오류 검출의 세 가지 조건을 만족하도록 설계되어진다. 제안한 방법의 주요 특징으로는 다음과 같다.

1) 데이터 은닉 전,후의 DNA 서열의 길이는 동일하며, 엑손의 DNA 정보는 데이터 은닉에 상관없이 변경되지 않는다.
2) 데이터 검출 과정에서는 은닉 키에 의하여 원 DNA 서열이 없이 데이터를 검출한다.
3) 뉴클레오티드 염기들의 이진 표현 방법이 쉽게 예측되지 않도록 한다.
4) 전송 및 변이에 의하여 발생되는 정보의 오류는 패리티 염기에 의하여 검출된다. 여기서 한 섹터 내에 하나의 염기 오류가 검출되며, 일부 조건을 만족할 경우 다중 염기 오류가 검출될 수 있다.

다음 절에서는 비부호 DNA 영역 상에서 데이터가 은닉 및 검출되는 과정들을 차례로 살펴보기로 한다. DNA 서열은 문자열이며, 은닉 및 추출 과정에서는 문자열을 이진 또는 정수 형태의 숫자열로 변경하여 사용되어진다. 본 논문에서는 문자열을 로마체로 숫자열을 이탤릭체로 구분하여 표현한다.

1. 메시지 은닉 과정

본 논문에서는 DNA 서열 S를 메시지 은닉 매개체라 한다. 제안한 방법에서는 그림 2(a)에서와 같은 과정에 의하여 DNA 서열 S를 인트론 (비부호 DNA)과 엑손 (부호 DNA)로 분리한 후 모든 인트론 내에 메시지를 은닉한다. 즉, 메시지는 인트론 섹터 단위로 은닉된다. 은닉될 이진 메시지가 $\boldsymbol{M}_b$이라 할 때, $\boldsymbol{MR}_b$의 은닉 과정은 다음과 같다.

1) 뉴클레오티드 염기 x의 이진 부호 x_b와 정수 부호 x 간의 변환 테이블 T를 생성한다. T는 데이터 은닉 및 검출을 위한 키로 사용된다. x, x_b, 및 x 간의 상관관계는 아래 식에 의하여 표현되어진다.

$$x = f(\mathrm{x});\ \mathrm{x} = f^{-1}(x), \tag{1}$$

$$x = h(x_b), x_b = h^{-1}(x), \tag{2}$$

$$\text{for } \mathrm{x} \in \{'\mathrm{A}', '\mathrm{C}', '\mathrm{G}', '\mathrm{T}'\}, x \in \{1,2,3,4(\text{or } 0)\}$$
$$x_b \in \{00,01,10,11\}$$

$f(\mathrm{x})$는 문자 변수의 염기 x를 정수 x로 변환하고, $h(x_b)$는 2비트의 이진 정보 x_b를 정수 x로 변환한다. 이들의 역함수는 각각 $f^{-1}(x) = \mathrm{x}$와 $h^{-1}(x) = x_b$와 같다. 입력되는 염기 x와 은닉 과정에서 수행되는 이진 정보 x_b와의 관계는

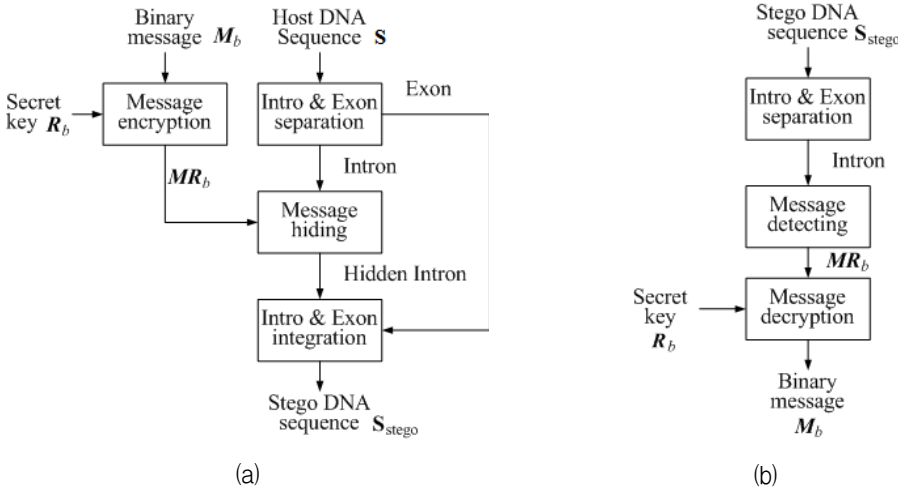

그림 2. 제안한 DNA 메시지 (a) 은닉 과정과 (b) 추출 과정
Fig. 2. Proposed DNA message (a) embedding process and (b) extracting process.

$$x_b = h^{-1}(f(\mathrm{x})),\ \mathrm{x} = f^{-1}(h(x_b)) \quad (3)$$

와 같이 표현된다.

2) 메시지의 보안성을 위하여 메시지 정보는 CS-PRBG(Cryptographically Secure Pseudo-random Bit Generator)[9], Blum-Blum-Shub 비트 생성기[10~11] 등에 의하여 생성된 랜덤 정보와 XOR 연산을 수행하여 암호화될 수 있다. 제안한 방법에서는 시드 $seed_R$의 CS-PRBG에 의하여 길이 L_R의 랜덤 이진 서열 $\boldsymbol{R}_b$를 생성한 다음, $\boldsymbol{M}_b$과 $\boldsymbol{R}_b$의 XOR 연산자에 의하여 암호화한 후, 오류정정을 위한 부호화율 1/3의 터보코드를 그림 3에서와 같이 수행한다. 따라서 최종 생성된 이진 메시지 $\boldsymbol{MR}_b$을 구한다.

$$\boldsymbol{MR}_b = \boldsymbol{MR}_x | \boldsymbol{MR}_{y1} | \boldsymbol{MR}_{y2} \quad (4)$$

where $\boldsymbol{MR}_x = \boldsymbol{M}_b \oplus \boldsymbol{R}_b$

$\boldsymbol{MR}_b$는 $h(\cdot)$에 의하여 정수 메시지 $\boldsymbol{MR}$으로 변환된다.

그림 3에서 복호기 DEC1의 복호화된 비트 $\boldsymbol{MR}_b(k)$에 대한 로그-우도비 (Logarithm of likelihood ratio, LLR) $\Lambda_1(\boldsymbol{MR}_b(k))$는

$$\Lambda_1(\boldsymbol{MR}_b(k)) = \log \frac{\Pr\{\boldsymbol{MR}_b(k) = 1 | observation\}}{\Pr\{\boldsymbol{MR}_b(k) = 0 | observation\}} \quad (5)$$

와 같이 정의된다. $\Pr\{\boldsymbol{MR}_b(k) = i | observation\}$, $i = 0, 1$는 $\boldsymbol{MR}_b(k)$의 사후 확률(A posteriori

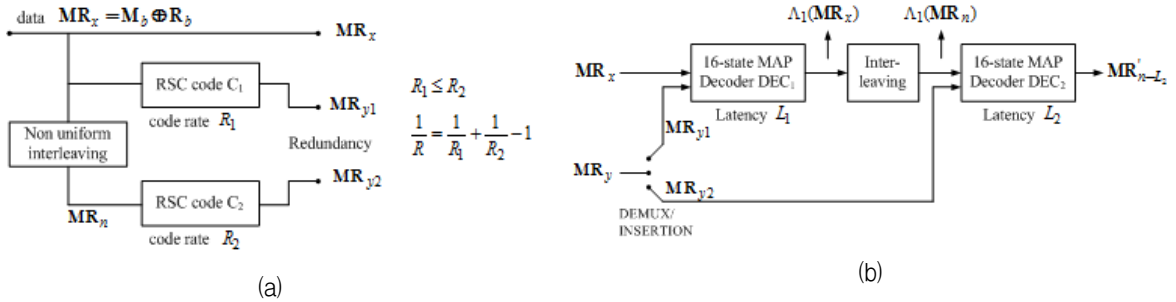

그림 3. 터보부호(부호화율=1/3)에 의한 (a) 메시지 생성 및 (b) 메시지 복호
Fig. 3. (a) Message generation and (b) message decoding by turbo code (coding rate=1/3).

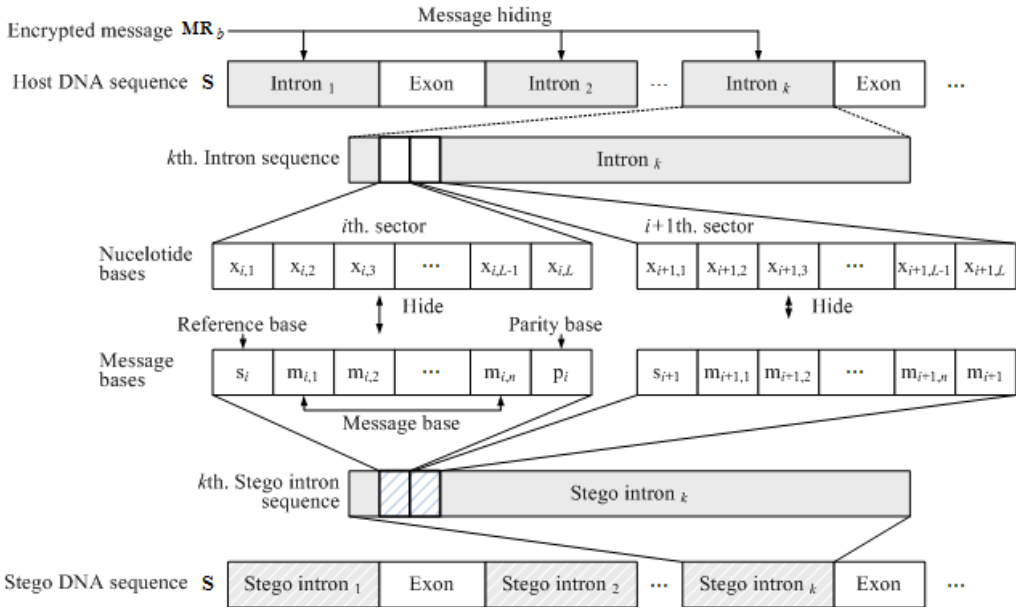

그림 4. DNA 서열 S 구조 및 임의의 인트론 서열 상에 섹터 단위로 은닉할 뉴클레오티드 염기와 메시지 구조
Fig. 4. Structure of DNA sequence S and nucleotide bases and messages for hiding in each sector of intron sequence.

probability, APP)이다.

3) DNA 서열 S의 구조는 그림 4에서와 같다. i번째 섹터의 길이가 L일 때, 메시지 염기 길이 n은 $n=L-2$와 같다. 즉, $L=n+2$이다. $x_{i,j}$는 i번째 섹터 내에 $j(j=1,2,\cdots,L)$번째 뉴클레오티드 염기라 한다.

첫 번째 염기 $x_{i,1}$는 에 대한 기준 염기 s_i는 섹터 내의 모든 염기 정수값들의 평균에 해당되는 염기로

$$x'_{i,1}=s_i \tag{6}$$

$$\text{where } s_i=f^{-1}(\lfloor \sum_{j=1}^{L} f(x_{i,j})/L+\frac{1}{2} \rfloor)$$

와 같이 사용된다. n개의 메시지 염기들은 n 비트의 암호화된 메시지가 은닉된 염기들로

$$x'_{i,k+1}=m_{i,k} \text{ for } k\in[1,n] \tag{7}$$

와 같이 사용되며, 마지막 번째 염기 $x_{i,L}$는 패리티 염기 p_i로

$$x'_{i,L}=p_i \tag{8}$$

와 같이 사용된다.

4) n 메시지 염기들은 기준값 $f(s_i)$와 암호화된 정수 메시지 $\boldsymbol{MR}$에 의하여 다음과 같이 부호된다.

$$m_{i,k}=(f(s_i)+\boldsymbol{MR}(k))\bmod(n+1) \tag{9}$$

$$m_{i,k}=f^{-1}(m_{i,k}) \tag{10}$$

섹터 상의 기준 문자는 {'A','C','T','G'} 중의 하나가 선택되며, 기준 문자는 섹터 마다 다르게 선택된다. 각 메시지는 각 섹터의 기준 문자에 따라 부호되므로, 섹터마다 랜덤한 이진 정보로 할당된다. 예를 들어, 메시지 문자 'A'는 기준 문자 'T'를 가지는 섹터 내에서는 '11'로 부호되며, 기준 문자 'G'를 가지는 섹터 내에서는 '10'으로 부호된다.

섹터 길이 변수 n은 용량성과 오류 가능성을 고려하여 결정된다. n의 값이 클수록 메시지를 은닉하기 위하여 필요한 뉴클레오티드 염기 수가 작아진다.

5) 패리티 체크 p_i와 패리티 염기 p_i는 각 섹터 상에서 패리티 염기 전의 모든 문자 부호들의 합에 의하여 계산된다.

$$p_i=(\sum_{j=1}^{L-1} f(x'_{i,j}))\bmod(4) \tag{11}$$

$$p_i=f^{-1}(p_i) \tag{12}$$

4) 단계의 예제에서 패리티 체크 p_i와 패리티 염기 p_i는 위의 수식에 의하여 다음과 같이 계산된다.

메시지 검출 과정에서는 이와 같은 패리티 염기 검출에 의하여 섹터 내에 오류 발생 여부를 확인한다. 제안한 방법에서는 한 섹터 내에 단일 염기 오류를 검출할 수 있으며, 특히 에러의 합이 0 (mod 4)이 아닐 경우 이중 염기 오류를 검출할 수 있다. 한 섹터 내에 염기가 많을수록 합이 0(mod 4)인 다중 오류 가능성이 커진다.

6) 모든 섹터 내에 메시지가 은닉될 때까지 4), 5) 단계를 반복적으로 수행한다.

S_{stego}는 공개 채널을 통하여 수신자에게 전송되며, 정수 변환표 T, PRBG의 시드 $seed_R$ 및 n은 비밀 채널을 통하여 수신자에게 전송된다.

다음은 비부호 DNA 서열이 S="GAACT…"이고, 암호화된 메시지가 $\boldsymbol{MR}=h(\boldsymbol{MR}_b)=123032\cdots$이고, $x\in\{'A','C','G','T'\}x\in\{0,1,2,3\}$이고, n이 3일 때 메시지 부호 과정의 예를 보여준다.

- 입력 서열 : S="GAACTCATAG…"
- 암호화된 정수 메시지

$$\boldsymbol{MR}_b=01\ 10\ 11\ 00\ 11\ 10$$

$$\boldsymbol{MR}=h(\boldsymbol{MR}_b)=123032\cdots$$

- 첫 번째 섹터 "GAACT"에 메시지 "123" 은닉
  - 기준 염기

$$s_1=f^{-1}(\lfloor \frac{f('G')+\cdots+f('T')}{5}+\frac{1}{2} \rfloor)$$
$$=f^{-1}(\lfloor \frac{2+0+0+1+3}{5}+\frac{1}{2} \rfloor)$$
$$=f^{-1}(1)='C'=x'_{1,1}$$

  - 메시지 염기

$$m_{1,1}=(f('C')+\boldsymbol{MR}(1))\bmod(4)=2\ ,$$
$$m_{1,1}=f^{-1}(m_{1,1})='G'=x'_{1,2}$$
$$m_{1,2}=(f('C')+\boldsymbol{MR}(2))\bmod(4)=3,$$
$$m_{1,2}=f^{-1}(m_{1,2})='T'=x'_{1,3}$$

$m_{1,3} = (f('C') + \boldsymbol{MR}(3)) \bmod (4) = 0,$

$\mathrm{m}_{1,3} = \mathrm{f}^{-1}(\mathrm{m}_{1,3}) = 'A' = \mathrm{x}'_{1,4}$

- 패리티 염기

$p_i = (\sum_{j=1}^{4} f(\mathrm{x}'_{i,j})) \bmod (4)$

$= (1+2+3+0) \bmod (4) = 2$

$\mathrm{p}_1 = \mathrm{f}^{-1}(\mathrm{p}_1) = 'G' = \mathrm{x}'_{1,5}$

· 최종 스테고 서열 S_{stego}="CGTAG…", (n=3)

2. 메시지 검출 과정

S_{stego}에 은닉된 메시지는 비밀 채널로 전송된 정수 변환표 T, PRBG의 시드 $seed_R$ 및 n에 의하여 검출된다. 검출 과정은 그림 2(b)에서와 같이 전송된 DNA 서열 S_{stego} 내에 인트론과 엑손을 분리한 후, 인트론 내에 은닉된 메시지를 차례로 검출한다. 검출 과정은 다음과 같다.

1) n를 알고 있으므로, 각 섹터 길이 $L(n+2)$를 구한 다음, 각 섹터의 첫 번째 문자는 기준 문자로, 마지막 문자는 패리티 체크로 사용한다.

2) 섹터 상에 오류를 검출하기 위하여 먼저 패리티 염기 p_i를 확인한다. 섹터 내에 오류가 발생하지 않을 경우 페리티 염기 p_i는 은닉 전의 염기와 동일하다. 이 때 메시지는 쉽게 검출될 수 있다. 그러나 패리티 염기가 같지 않다면, 메시지의 전송 오류가 발생한 것이다.

3) 염기 'T'를 기준 문자로 사용하여 메시지를 복호한다. 이 때 다음과 같은 수식에 의하여 $m_{i,k}$과 S_i의 차이가 계산된다.

$$\boldsymbol{MR}(k) = (f(\mathrm{m}_{i.k}) - f(\mathrm{s}_i)) \bmod (n+1) \tag{13}$$

$$\boldsymbol{MR}_b(k) = h^{-1}(\boldsymbol{MR}(k)) \tag{14}$$

4) 모든 섹터들의 메시지를 복원하기 위하여 2), 3)의 단계를 반복 수행한다. 수행이 종료될 경우 $\boldsymbol{MR}_b$이 생성된다.

5) 시드 $seed_R$를 사용하여 유사 랜덤 이진 정보 $\boldsymbol{R}_b$을 생성한 다음, $\boldsymbol{R}_b$과 $\boldsymbol{MR}_b$를 XOR으로 수행하여 복원 메시지 $\hat{\boldsymbol{M}}_b$를 생성한다.

전송된 DNA 서열이 $\mathrm{S}_{\mathrm{stego}}$="CGTAG……"이고 n=3일 때, 이진 메시지 $\boldsymbol{MR}_b$은 다음과 같이 계산되어진다.

- $\boldsymbol{MR}(1) = (f('G') - f('C')) \bmod (4) = 1,$
  $\boldsymbol{MR}_b(1) = h^{-1}(1) = 01$
- $\boldsymbol{MR}(2) = (f('T') - f('C')) \bmod (4) = 2,$
  $\boldsymbol{MR}_b(2) = h^{-1}(2) = 10$
- $\boldsymbol{MR}(3) = (f('A') - f('C')) \bmod (4) = 3,$
  $\boldsymbol{MR}_b(3) = h^{-1}(3) = 11$
- $\boldsymbol{MR}_b$="01 10 11…"

$\boldsymbol{MR}_b$ 상에서 터보 복호기로부터 $\boldsymbol{MR}_x = \boldsymbol{M}_b \oplus \boldsymbol{R}_b$를 추출한 후, $\boldsymbol{R}_b$에 의하여 $\boldsymbol{M}_b$가 얻어진다. 이상으로부터 오류없이 검출 과정이 끝날 경우, 복원 이진 정보 $\hat{\boldsymbol{M}}_b$를 원본 메시지로 번역된다.

Ⅳ. 실험 결과

제안한 방법의 평가를 위한 실험에서는 표 1에서와 같이 NCBI 데이터베이스[12]로부터 4개의 DNA 서열을 사용하였으며, MATLAB 툴을 이용하여 in silico 기반으로 수행되었다. 본 실험에서는 제안한 방법의 섹터 길이 변수 n를 8으로 하여, 한 섹터 길이가 10 염기가 되도록 하였다.

본 실험 평가에서는 비부호 영역 DNA 워터마킹 방법인 Clelland[3], Leier[4], Heider[5], Shiu[6] 방법과의 데이터 용량 및 기능에 대하여 비교하였다. 여기서 부호 영역 DNA 워터마킹[7~8]은 코돈 기반 단백질 보존 조건에 해당되는 방법으로, 비교 평가에서 제외되었다.

1. 데이터 용량

본 실험에서는 제안한 방법의 데이터 용량 성능 평가를 위하여 bpn를 계산하여 이를 비교하였다. bpn(bit per nucleotide base)은 주어진 DNA 서열 S상에 최대 은닉되는 염기 당 비트수를 나타내는 것으로,

$$\mathrm{bpn} = \frac{|\mathrm{W}|}{|\mathrm{S}| + |\mathrm{P}|} [\mathrm{bit}/\triangle] \tag{15}$$

와 같이 정의되어진다. 여기서 |W|는 은닉되는 메시지

의 비트 수를 나타내고, |S|는 DNA 서열의 총 염기 수를 나타낸다. 그리고 |P|는 데이터 은닉에 필요한 부가적인 페이로드(payload) 염기 수를 나타낸다. 즉 |S|+|P|는 |W| 비트 은닉에 필요한 총 염기 수를 의미한다.

제안한 방법은 한 섹터 내에 n개의 메시지 염기 m들과 기준 염기 s, 패리티 염기 p들로 구성되며, 각 염기 당 2비트의 정보가 은닉되며, 부가적인 페이로드 염기가 필요없다. 한 섹터의 길이가 $L=n+2$이므로 제안한 방법의 bpn은

$$\begin{aligned} \mathrm{bpn}_{Proposed} &= \frac{|\mathbf{W}|}{|\mathbf{S}|} \\ &= \frac{\lfloor \frac{|\mathbf{S}|}{n+2} \times 2n \rfloor}{|\mathbf{S}|} [\mathrm{bit/base}] \end{aligned} \quad (16)$$

와 같다. 본 실험에서는 섹터 길이 L=10 (n=8)이므로 $\mathrm{bpn}_{Proposed}$=1.6이다. 기준 염기와 패리티 염기에 해당되는 비트들은 |W|에서 제외되나, 에러 검출에서 매우 필요한 정보들이다.

Clelland 방법[3]은 64개의 문자(6비트)들로 구성된 데이터들을 하나의 문자 당 3개 염기에 은닉하는 것으로 시작 마커와 종료 마커 및 은닉된 염기들을 DNA 서열에 추가한다. 즉, 기존 DNA S의 변경없이 페이로드 염기 |P|가 은닉된 염기이므로, bpn은

$$\mathrm{bpn}_{Clelland} = \frac{(|\mathbf{P}|-6)\times 2}{|\mathbf{S}|+|\mathbf{P}|} [\mathrm{bit/base}] \quad (17)$$

와 같다. 제안한 방법과 동일한 데이터 비트수와 동일하게 놓을 경우 $\mathrm{bpn}_{Clelland}$ =0.889이다. Leier 방법[4]은 이진 정보를 나타내는 짧은 서열을 생성하여 이를 시작 마

표 1. 실험에 사용된 DNA 서열과 대상 염기 개수, 페이로드, 및 bpn 결과
(Base : 데이터 비트와 염기 개수가 주어졌을 때, 은닉에 필요한 염기 수)
Table 1. Results of target bases, payload, and bpn for test DNA sequence.

DNA 서열	Species 정의	Number of nucleotide bases	Category	Proposed method	Clelland [3]	Leier [4]	Heider [5]	Shiu [6]		
								Insertion	Complementary	Substitution
GJ060459	Bos taurus chromosome 1	213,382	Base	152,416	365,804	1,737,542	213,381	365,798	3,693,382	213,381
			Payload	0	152,422	1,524,160	0	152,416	3,480,000	0
			bit	304,832	304,832	304,832	142,254	304,832	304,832	142,254
			bpn	1.428	0.833	0.175	0.667	0.833	0.082	0.667
GJ061116	Bos taurus chromosome 7	187,446	Base	133,890	321,342	1,526,346	187,446	321,336	3,427,446	187,446
			Payload	0	133,896	1,338,900	0	133,890	3,240,000	0
			bit	267,780	267,780	267,780	124,964	267,780	267,780	124,964
			bpn	1.428	0.833	0.175	0.667	0.833	0.078	0.667
GJ061125	Bos taurus chromosome 7	221,140	Base	157,958	379,104	1,800,720	221,139	379,098	3,701,140	221,139
			Payload	0	157,964	1,579,580	0	157,958	3,480,000	0
			bit	315,916	315,916	315,916	147,426	315,916	315,916	147,426
			bpn	1.428	0.833	0.175	0.667	0.833	0.085	0.667
NW_001502389	Bos taurus breed Hereford chromosome 19	171,877	Base	122,771	294,654	1,399,587	171,876	294,648	3,411,877	171,876
			Payload	0	122,777	1,227,710	0	122,771	3,240,000	0
			bit	245,542	245,542	245,542	114,584	245,542	245,542	114,584
			bpn	1.428	0.833	0.175	0.667	0.833	0.072	0.667

표 2. 제안한 방법과 기존 방법들 간의 주요 특징 비교
Table 2. Comparison of main features between our method and related methods.

Method	Store Digital Data?	Error Handling	Location of Data	Preserve Amino Acid?	Original DNA required?
Proposed method	yes	yes	intron	yes	no
Clelland [3] : DNA microdot	no	no	intron	no	no
Leier [4] : DNA binary strand	yes	no	intron	no	no
Heider [5] : DNA crypt	yes	yes	intron & exon	yes	no
Shiu [6] : Insertion method	yes	no	intron	yes	no
Shiu [6] : Complementary pairs method	yes	no	intron	yes	no
Shiu [6] : Substitution method	yes	yes	intron	yes	yes

커와 종료 마커에 대한 키 서열을 연결한다. 따라서 하나의 비트는 n길이의 짧은 서열과 키 서열로 구성된 $n+2$개의 염기 서열이 기존 서열에 추가된다. 그러므로 이 방법의 bpn은

$$\mathrm{bpn}_{\mathrm{Leier}} = \frac{\lfloor |\mathbf{P}|/(n+2) \rfloor}{|\mathbf{S}|+|\mathbf{P}|} [\mathrm{bit}/\triangle] \quad (18)$$

와 같이, 다른 방법에 비하여 bpn이 매우 낮다. 본 실험에서는 최소 길이 n=3으로 설정하였다. Heider 방법[5]은 페이로드 염기 필요없이 에러 정정을 위하여 m번 반복하여 은닉하는 것으로, bpn은

$$\mathrm{bpn}_{\mathrm{Heider}} = \frac{\lfloor |\mathbf{S}|/m \rfloor \times 2}{|\mathbf{S}|} [\mathrm{bit}/\triangle] \quad (19)$$

와 같다. Shiu[6]의 방법들 중 삽입 방법은 Clelland의 시작 및 종료 마커와 제외한 bpn과 유사하며, 상보형 쌍 방법은 DNA 사슬에서 가장 긴 서열이 될 유일한 염기 서열 $\mathbf{S}_{\max} = \mathbf{P}$이 추가적으로 필요하다. 따라서 이 방법은 bpn이 매우 낮다. 마지막으로 치환 방법은 염기당 최대 2비트 삽입 가능하며, 이를 m번 반복하여 은닉된다. Shiu의 세 방법에 대한 bpn은 다음과 같다.

$$\mathrm{bpn}_{\mathrm{Insertion}} = \frac{2|\mathbf{P}|}{|\mathbf{S}|+|\mathbf{P}|} [\mathrm{bit}/\triangle] \quad (20)$$

$$\mathrm{bpn}_{\mathrm{Complementary}} = \frac{2|\mathbf{S}|_{\max}}{|\mathbf{S}|+|\mathbf{S}|_{\max}} [\mathrm{bit}/\triangle] \quad (21)$$

$$\mathrm{bpn}_{\mathrm{Substitution}} = \frac{|\mathbf{S}| \times 2/m}{|\mathbf{S}|} [\mathrm{bit}/\triangle] \quad (22)$$

제안한 방법의 섹터 길이 L=10일 때, $\mathrm{bpn}_{Proposed}$=1.428이다. 제안한 방법의 동일한 데이터 비트 수에 대하여 Clelland 방법은 $\mathrm{bpn}_{\mathrm{Clelland}}$=0.833이고, Leier 방법은 $\mathrm{bpn}_{\mathrm{Leier}}$=0.175이다. Shiu 방법의 삽입, 상보 쌍 방법은 $\mathrm{bpn}_{\mathrm{Insertion}}$=0.8333, $\mathrm{bpn}_{\mathrm{Complementary}}$=0.08이다. Shiu의 치환 방법과 Heider 방법은 페이로드없이 반복 은닉으로 인하여 동일한 비트 수를 설정할 수 없으므로, 최소 반복 횟수 m=3으로 놓을 경우, bpn_{Heider}=0.667이고, $\mathrm{bpn}_{Substitution}$=0.667로 동일하다. 이상의 제안한 방법과 기존 방법들의 bpn에 대한 평가 결과는 표 1에서와 같다. 표 상에서 Bases는 염기 서열과 데이터 비트가 주어졌을 때 삽입에 필요한 염기 수를 나타낸다. 제안한 방법의 bpn이 기존 방법들보다 약 0.59에서 1.34 정도 높음을 알 수 있었다.

2. 기존 방법과의 비교 평가

제안한 방법과 기존의 6가지 방법들 간의 주요 특징에 대한 비교를 표 2에 나타내었다. 초창기 Clelland 등[3]과 Leier 등[4]의 방법들은 오류 정정 및 아미노산 보존성이 가능하지 못하다. Heider 등[5]의 방법은 오류 정정과 아미노산 보존성이 가능하며, 부호 영역인 exon 영역 및 비부호 영역인 intron 영역에 적용이 가능하다. 그러나 반복 횟수에 의하여 데이터 용량이 다소 제한적이다. 최근 Shiu 등[6]의 삽입 및 상보형 쌍 방법들은 비부호 영역에 은닉하므로 아미노산 보존이 가능하나, 오류 정정이 가능하지 못하며, 데이터 용량도 크지 않다. 그러나 Shiu 등의 치환 방법은 오류 정정이 가능하나, 데이터 추출시 원본 DNA 서열이 필요하다. 이와 반면 제안한 방법은 데이터 저장, 오류 검출 및 정정, 아미노산 보존, 및 원본 DNA 서열이 필요없는 블라인드 검출 등의 특징을 모두 가진다.

V. 결 론

DNA은 대용량의 디지털 정보를 장시간 저장할 수 있는 새로운 저장 매개물로 인식되고 있다. 본 논문에서는 DNA 대용량 데이터 저장을 위하여 단백질 보존성, 원 DNA 서열없이 데이터 추출성, 오류 검출 및 정정 능력을 가지는 DNA 데이터 은닉 방법을 제안하였다. 제안한 방법에서는 단백질 보존을 위하여 비부호 영역에 데이터를 은닉하고, 섹터 단위별 패러티 염기를 이용하여 데이터 오류 검출한 다음, Reed Solomon 부호를 이용한 퍼지 논리 제어를 이용하여 오류를 정정한다. 실험 결과로부터 제안한 방법이 기존 방법들에 비하여 높은 데이터 용량과 오류 정정 능력을 가짐을 확인하였다.

REFERENCES

[1] P. C. Wong, K. K. Wong, and H. Foote, "Organic data memory using the DNA approach," *Communications of the ACM*, vol. 46,

issue 1, pp. 95-98, Jan. 2003.
[2] G. Church, Y. Gao, and S. Kosuri, "Next-Generation Digital Information Storage in DNA," *Science*, vol. 337, no. 6102, pp. 1628, Sept. 2012.
[3] C. Clelland, V. Risca, and C. Bancroft, "Hiding messages in DNA microdots," *Nature*, vol. 399, pp. 533-534, June 1999.
[4] A. Leier, C. Richter, W. Banzhaf, and H. Rauhe, "Cryptography with DNA binary strands," *BioSystems*, vol. 57, issue 1, pp. 13-22, June 2000.
[5] D. Heider and A. Barnekow, "DNA-based watermarks using the DNA-Crypt algorithm," *BMC Bioinformatics*, vol. 8, no. 176, May 2007.
[6] H. Shiu, K. Ng, J. Fang, R. Lee, and C. Huang, "Data hiding methods based upon DNA sequences," *Information Sciences*, vol. 180, issue 11, pp. 2196 - 2208, June 2010.
[7] S.-H. Lee, S.-G. Kwin, and K.-R. Kwon, "Robust DNA Watermarking based on Coding DNA Sequence," *Journal of the Institute of Electronics and Information Engineers*, vol. 49, issue 2, pp. 123-133, March 2012.
[8] S.-H. Lee, S.-G. Kwin, and K.-R. Kwon, "A Robust DNA Watermarking in Lifting Based 1D DWT Domain," *Journal of the Institute of Electronics and Information Engineers,* vol. 49, issue 10, pp. 91-101, October 2012.
[9] A. Menezes, P. van Oorschot, and S. Vanstone, *Handbook of Applied Cryptography*, CRC Press, pp. 169-190, 1996.
[10] A. Sidorenko and B. Schoenmakers, "Concrete Security of the Blum-Blum-Shub Pseudorandom Generator," Cryptography and Coding: 10th IMA International Conference, LNCS, vol. 2796, pp. 355-375, Dec. 2005.
[11] P. Junod, "Cryptographic Secure Pseudo-Random Bits Generation : The Blum-Blum-Shub Generator," http://crypto.junod.info/bbs.pdf, 1999.
[12] NCBI, National Center for Biotechnology Information, http://www.ncbi.nlm.nih.gov., Accessed on Jan. 2014.

저 자 소 개

이 석 환(정회원)
1999년 경북대학교 전자공학과 학사 졸업.
2001년 경북대학교 전자공학과 석사 졸업.
2004년 경북대학교 전자공학과 박사 졸업.
2005년～현재 동명대학교 정보보호학과 부교수
<주관심분야 : 워터마킹, DRM, 영상신호처리, 3D 그래픽스>

권 기 룡(정회원)
1986년 경북대학교 전자공학과 학사 졸업.
1990년 경북대학교 전자공학과 석사 졸업.
1994년 경북대학교 전자공학과 박사 졸업.
2000년～2001년 Univ. of Minnesota, Post-Doc.
1996년～2006년 부산외국어대학교 컴퓨터전자공학부 부교수
2006년～현재 부경대학교 IT융합응용공학과 교수
<주관심분야 : 멀티미디어 정보보호, 멀티미디어 통신 및 신호처리>