Journal of Korea Multimedia Society Vol. 20, No. 2, February 2017(pp. 244-253)
https://doi.org/10.9717/kmms.2017.20.2.244

바이오 정보보호 위한 히스토그램 쉬프팅 기반 가역성 DNA 워터마킹 기법

이석환[†], 권성근[††], 이응주[†††], 권기룡[††††]

Reversible DNA Watermarking Technique Using Histogram Shifting for Bio-Security

Suk-Hwan Lee[†], Seong-Geun Kwon[††], Eung-Joo Lee[†††], Ki-Ryong Kwon[††††]

ABSTRACT

Reversible DNA watermarking is capable of continuous DNA storage and forgery prevention, and has the advantage of being able to analyze biological mutation processes by external watermarking by iterative process of concealment and restoration. In this paper, we propose a reversible DNA watermarking method based on histogram multiple shifting of noncoding DNA sequence that can prevent false start codon, maintain original sequence length, maintain high watermark capacity without biologic mutation. The proposed method transforms the non-coding region DNA sequence to the n-th code coefficients and embeds the multiple bits of the n-th code coefficients by the non-recursive histogram multiple shifting method. The multi-bit embedding process prevents the false start codon generation through comparison search between adjacent concealed nucleotide sequences. From the experimental results, it was confirmed that the proposed method has higher watermark capacity of 0.004-0.382 bpn than the conventional method and has higher watermark capacity than the additional data. Also, it was confirmed that false start codon was not generated unlike the conventional method.

Key words: Reversible DNA Watermarking, DNA Security, Histogram Shifting, DNA Information Hiding

1. 서 론

DNA 시퀀싱(sequencing) 및 DNA 컴퓨팅 기술의 비약적인 발달로 인하여 개인 유전정보에 대한 프라이버시 침해가 발생될 것이다. 개인 유전정보 보호를 위하여 유전정보의 수집절차 적법성 및 공정성과 제도적인 안정망이 구축되어야 하며, HGI(human genome information) 사용에 대한 윤리적인 법 이외에도 불법 복제 및 도용 방지 및 인증을 위한 기술이 필요하다. 또한 DNA는 1g 당 약 10^8 Tera byte를 저장할 수 있는 대용량 생체 정보 매개체로 인식됨에 따라 새로운 DNA 암호화 시스템에 대한 필요성이

※ Corresponding Author : Suk-Hwan Lee, Address: (48520) Shinseonn-Ro 428, Namgu, Busau, Korea, TEL : +82-51-629-6257, FAX : +82-51-629-1129, E-mail : skylee@tu.ac.kr
Receipt date : Aug. 9, 2016, Revision date : Dec. 14, 2016
Approval date : Jan. 7, 2017
[†] Dept. of Information Security, Tongmyong University
[††] Dept. of Electronics Engineering, KyungIl University (E-mail : sgkwon@kiu.ac.kr)
[†††] Dept. of Information Communication Engineering, Tongmyong University (E-mail : ejlee@tu.ac.kr)
[††††] Dept. of IT Convergence and Application Engineering, Pukyong National University (E-mail : krkwon@pknu.ac.kr)
※ This research was supported by supported by BB21 (Brain Busan) Project and Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2016R1D1A3B03931003, NRF-2014R1A1A4A01006663 and NRF-2011-0023118)

제기되고 있다. 이와 같은 필요성에 따라 최근 DNA 서열에 대한 암호화[1], 비밀 통신위한 DNA 스테가노그라픽[2], 및 GMO(genetically modified organism) 또는 서열의 인증 및 저작권 보호위한 DNA 워터마킹 기술[3]-[5]에 대한 연구가 이루어진다.

기존 멀티미디어와는 달리 DNA 정보은닉 기술에서는 워터마크에 의한 생물학적 변이가 전혀 없어야 하며, 손실없이 원본 DNA 서열이 복원되어야 한다. 대부분의 DNA 스테가노그라픽 및 워터마킹 기법[2-5]들은 원본 DNA 서열의 복원 기능이 없는 비가역에 해당되므로, 가역 DNA 스테가노그라픽 및 워터마킹에 대한 연구가 매우 필요하다. 특히 가역 DNA 워터마킹은 연속적인 DNA 저장과 위변조 방지가 가능하며, 은닉과 복원의 반복 과정에 의하여 외부 워터마크에 의한 생물학적 변이 과정을 분석할 수 있는 장점을 가진다.

가역 DNA 워터마킹은 높은 용량성 외에 허위개시코돈(false start codon)이 발생되지 않아야 한다. DNA은 부호 영역과 비부호 영역으로 나뉘며, 부호 영역은 생물학적 기능을 가지는 아미노산으로 변환되는 반면, 비부호 영역은 아미노산으로 변환되지 않는다. DNA 서열에서 부호 영역은 “ATG”의 개시코돈(start codon)으로 시작되며, 이 코돈(codon)은 mRNA과 리보솜과 결합해 단백질 번역을 시작하도록 한다. 워터마크에 의하여 비부호 영역의 일부 염기서열이 개시코돈으로 바뀔 경우, 이 영역은 부호 영역으로 변경되어 DNA 변이가 발생된다. 본 논문에서는 이를 허위개시코돈이라 정의한다.

기존의 가역 DNA 워터마킹 방법들을 살펴보면, Chen 등[6]은 “ATCG”의 염기 서열들을 |w| 비트 단위의 십진수 부호서열로 변환한 다음, 인접한 부호 쌍에 속한 부호 쌍의 차이에 비밀메시지 비트가 은닉된다. 이 방법은 평균 0.09bpn～0.13bpn의 낮은 은닉 용량을 가지며, 허위개시코돈이 발생되는 단점을 가진다. Huang 등[7]은 십진수 부호서열에서 가장 높은 빈도수를 가지는 십진수 부호 p_j를 구한 다음, 은닉 비트가 1이면, p_j를 가장 낮은 빈도수의 부호로 변경하고, 그렇지 않으면 그대로 유지한다. 이 방법은 1.86%～4.07%의 낮은 염기 변화율을 가지나, 0.011bpn～0.024bpn의 매우 낮은 은닉 용량을 가지며, Chen의 방법과 같이 허위개시코돈의 발생된다. Lee 등[8,9]은 현재 부호계수에 대한 최소자승 예측오차를 구한 다음, 예측오차 확장 조건에 따라 결정된 비트수만큼 예측오차를 확장한다. 이때 은닉된 인접 염기서열 간의 비교탐색을 통하여 허위개시코돈 생성을 방지한다. 이 방법은 기존 방법의 문제점인 허위개시코돈을 방지하고, 0.413 bpn 은닉 용량을 가지나, 평균 1.788bpn 정도의 부가데이터가 필요하다. 이 방법들은 상보쌍 염기 치환 기반 데이터 은닉으로 워터마크 추출 및 복원시 참조되는 원본 DNA 서열이 필요하다. 이상의 기존 방법들은 원본 서열 길이를 유지하나, 허위개시코돈 방지를 고려하지 않고, 비블라인드이거나, 낮은 용량을 가진다.

가역성 영상 워터마킹에서는 높은 용량을 위하여 히스토그램 쉬프팅 이론[10,11] 기반으로 제안되어지고 있다. 이들 방법들은 은닉 대상의 최대점 주위의 히스토그램 빈(bin)을 쉬프팅하여 공간을 확보한 다음, 워터마크 비트에 따라 최대점을 좌우로 쉬프팅한다. 여기서 높은 워터마크 용량을 위하여 최대점들이 많이 확보되어야 하나, 화질 열화를 방지하기 위하여 쉬프팅 구간들이 제한되어야 한다. 영상의 화소 및 화질 속성과는 달리, DNA 서열은 염기로 구성되어 있으며, 허위개시코돈과 같은 생물학적 변이가 유발되지 않는 한, 쉬프팅 구간의 제한이 크지 않다. 본 논문에서는 허위개시코돈 방지 및 높은 용량성을 가지는 히스토그램 쉬프팅 기반 가역 DNA 워터마킹 방법들을 제시한다.

제안한 방법의 주요 특징으로는 높은 용량 및 DNA 서열의 신호 처리 용이성을 위하여 비부호 DNA 영역의 염기서열을 연속적인 n개 염기 단위의 정수열로 변환한다. DNA 정수열의 히스토그램 영역을 구간별로 구분한 다음, 경계 구간을 제외한 나머지 인접 구간 간의 히스토그램들은 워터마크 다중 비트에 따라 서로 쉬프팅을 한다. 은닉 과정에서는 워터마크된 염기서열 내에 인접 염기들 간의 비교 탐색을 위하여 개시코돈 생성을 방지한다. 제안한 방법과 기존의 Chen 방법[6]과 Huang 방법[7], 및 LS-PE 방법[8,9]과의 비교 실험에서 제안한 방법이 워터마크 용량과 용량 효율이 높음을 확인했다. 즉, 워터마크 용량성은 제안한 방법, LS-PE 방법, Chen 방법, Huang 방법으로 순으로 높게 나타났으며, 워터마크 용량 대비 부가데이터 용량의 비율에서 Chen 방법, 제안한 방법, Huang 방법, LS-PE 방법으로 순으로 높게 나타났다. 제안한 방법과 LS-PE 방법은 허위

개시코돈이 발생되지 않으나, Chen 방법과 Huang 방법은 약 10^4~5.78×10^5 염기 당 하나의 허위개시코돈이 발생됨을 확인했다.

본 논문의 구성은 다음과 같다. 2장에서는 가역 DNA 정보은닉 이론에 대하여 간략히 살펴보고, 3장에서는 제안한 히스토그램 쉬프팅 기반 가역 워터마킹 방법에 대하여 자세히 살펴본다. 4장에서는 제안한 방법과 기존 방법과의 비교 실험을 분석 한 후, 마지막 5장에서는 본 논문의 결론을 맺는다.

2. 가역 DNA 워터마킹 이론

가역 DNA 워터마킹은 원본 DNA 서열의 손실없이 복구가 가능한 것으로, 부호 DNA와 비부호 DNA에 대하여 적용이 가능하다. 부호 DNA에서는 단백질 부호 보존으로 인하여 중첩코돈(degeneracy codon) 내에 가역 워터마크가 은닉되어야 한다. 동일한 아미노산 내에서 워터마크에 의하여 변경될 수 있는 범위는 n중(n-fold) 중첩코돈 (n=1,2,3,4,6) 내에 n개 코돈뿐이다. 부호 DNA은 이와 같은 매우 작은 중첩코돈 범위로 인하여 가역 워터마킹에 적합하지 않다. 비부호 DNA은 단백질 부호 보존의 조건이 없으므로, 부호 DNA보다 워터마크 가용 범위는 다소 높다. 이러한 이유로 비부호 DNA는 DNA 가역 워터마크 은닉 대상으로 적합하다.

비부호 DNA에 대한 기본적인 가역 워터마킹의 기본적인 과정은 Fig. 1에서와 같다. 먼저 DNA 서열 상에서 비부호 영역의 염기 서열들은 정수열로 변환된 후, 오류 정정 또는 암호화된 워터마크가 가역성을 가지도록 정수열에 은닉된다. 그런 다음 워터마크된 정수열은 염기 서열로 변환된 후, 부호 영역과 연결되어 워터마크된 DNA 서열이 생성된다.

비부호 DNA은 멀티미디어 데이터의 화질과 같은 속성은 없으나, 고려해야 될 주요 특징이 있다.

1) 첫 번째로는 염기 서열은 네 가지 문자(A,T,C, G(or U))로 구성되므로, 멀티미디어 데이터에 비하여 매우 낮은 동적 범위를 가진다. 즉, 하나의 염기는 2비트의 정보를 가지며, 8비트의 영상 화소값에 비하여 매우 낮은 레벨을 가진다. 염기 서열에 대한 신호처리가 제한적이므로, 이로 인한 은닉 데이터의 용량이 낮다. 하지만, 연속된 염기 서열을 조합할 경우 다중 비트의 정보를 가지게 된다. 예를 들어, 4개의 연속된 염기 서열은 4^4=256 레벨을 가지게 되므로, 8비트 영상 화소값과 같은 레벨을 가진다.

2) 두 번째로는 은닉 과정에 의하여 비부호 DNA의 일부 염기 서열이 부호 DNA 서열의 개시코돈(Methionine, “ATG”)으로 변경될 가능성이 있다. 워터마크 은닉 과정에서 개시코돈으로 변경될 염기 서열들을 예측하여 제외하여야 한다.

3) 세 번째로는 DNA 서열의 길이가 변경되지 않으면서, 참조 서열 또는 원본 서열이 없이 워터마크 추출 또는 DNA 서열 복원이 가능하여야 한다(blind). 워터마크 정보를 담은 염기 서열들이 비부호 DNA 서열의 정크 영역에 추가(append) 및 추출될 수 있다. 외부 염기 서열 추가에 의한 생물학적 기능의 규명이 명확하지 않으므로, 주어진 DNA 서열 내에 워터마크가 은닉되어야 한다.

3. 제안한 히스토그램 쉬프팅 기반 가역 DNA 워터마킹 방법

DNA 서열은 Exon의 부호 영역과 Intro의 비부호 영역으로 나뉘어진다. Fig. 2는 임의의 DNA 서열 상에 가역 워터마킹의 예를 보여준다. 비부호 영역에

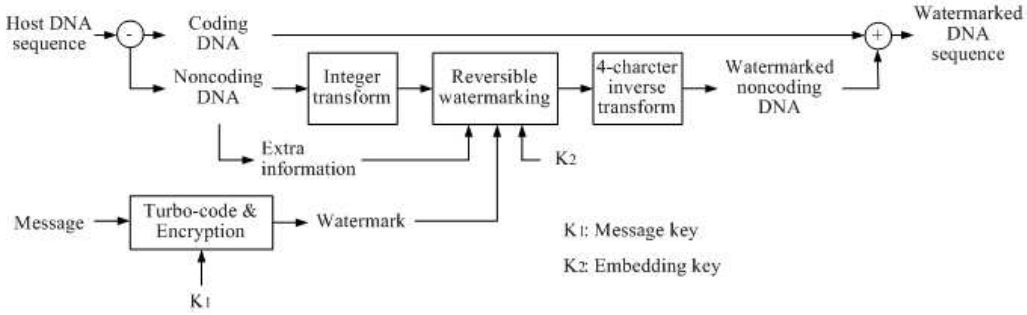

Fig. 1. General process of reversible watermarking for noncoding DNA sequence.

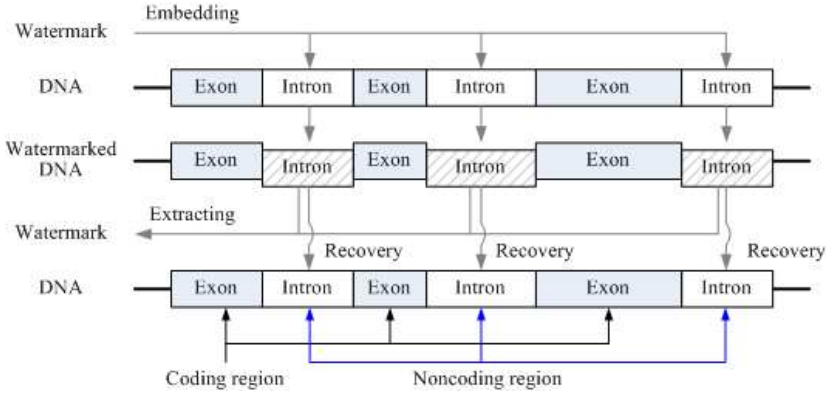

Fig. 2. An example of reversible DNA watermarking for a DNA sequence.

가역 워터마크가 은닉되며, 은닉된 비부호 영역들은 추출 및 복원 과정에 의하여 워터마크 추출 및 원본 영역으로 복원된다.

제안한 방법은 비부호 DNA 정수열에 대하여 히스토그램 쉬프팅 기반의 가역 워터마크를 은닉하는 것으로, 주요 특징은 다음과 같다. (1) 블라인드 가역성 (Blind Reversibility) : DNA 서열의 길이 변경 및 아미노산 변경없이 가역 워터마크가 은닉되며, 원본 DNA 서열없이 추출 및 복원이 가능하도록 한다. (2) 워터마킹 용이성 (Watermarking Usability) : 2비트 염기 서열을 $2n$비트의 부호계수 서열(code value sequence)로 부호화하여 가역 워터마크 은닉, 추출 및 복원 과정이 용이하도록 한다. (3) 워터마크 용량 (Watermark Capacity) : DNA 정수열의 히스토그램 쉬프팅 기반으로 다중 비트 은닉(multi-bit embedding)이 가능하도록 하여 워터마크 용량성을 높인다. (4) 허위개시코돈 방지 : 허위개시코돈 정수 테이블과 인접 부호계수 간 비교탐색을 통하여 부호계수 내 및 부호계수 간 허위개시코돈 생성을 방지한다.

본 장에서는 먼저 DNA 정수열의 다중비트 부호화에 대하여 살펴본 후, 이들 히스토그램 쉬프팅 기반 다중 비트 은닉, 복원, 및 추출 과정에 대하여 각각 살펴보기로 한다.

3.1 히스토그램 쉬프팅 기반 워터마크 부호 과정

3.1.1 4-문자 염기 정수화

뉴클레오티드 염기는 b=(A, T, C, G) 4-문자로 표현되며, 이를 4개의 십진수 또는 2비트의 이진수로 표현된다;

$b=(0,1,2,3)_{10}=(00,01,10,11)_2 \leftarrow \mathrm{b}=(\mathrm{A,T,C,G})$.

신호 처리 용이성을 위하여 2비트 계수보다 2비트 이상의 다중비트로 표현된 계수로 확장하여야 한다. 제안한 방법에서는 n개 염기들로 구성된 염기 블록 $\mathbf{x}$ 단위로 $2n$비트의 부호계수 x로

$$x=f(\mathbf{x})=\sum_{k=1}^{n}(b_k \cdot 2^{2(n-k)}) \text{ where } \mathbf{x}=(\mathrm{b}_1,\mathrm{b}_2,\cdots,\mathrm{b}_n),\ x\in[0,2^{2n}-1] \quad (1)$$

와 같이 부호한다. 정수 부호 x로부터 염기 블록의 염기들은

$$f^{-1}(x)=\mathbf{x} \text{ where } b_k=(x \gg 2(n-k))\%4 \text{ for } k=1,\cdots,n \quad (2)$$

와 같이 쉽게 복원된다. 본 논문에서는 염기 블록의 염기 개수 n를 부호차수로 부르기로 한다. 은닉 영역 D_k 내 염기들은 부호차수 n에 의하여 정수열 X_k로 부호된다; $\mathrm{X}_k=\{x_i|i\in[1,N_k]\}$. 이 때 N_k은 부호차수 n에 의하여 결정된다; $N_k=\lfloor|\mathrm{D}_k|/n\rfloor$.

3.1.2 은닉 영역 선택 (Target region)

임의의 DNA은 다양한 길이를 가지는 비부호 영역을 가진다. 이 때 염기 개수가 작은 영역들은 정수열의 길이가 짧아지므로, 워터마크 은닉 대상에 적합하지 않다. 제안한 방법에서는 α개 이상의 정수 부호를 가지는 영역을 선택하여, 부호차수 n에 대한 은닉 영역들의 집합 $\Gamma(n)$을

$$\Gamma(n)=\{\mathrm{D_i} \mid |\mathrm{D_i}|>\alpha\mathrm{p}\times\mathrm{n}\},\ \mathrm{D_i}=\{\mathrm{b_{ij}}|\mathrm{j}\in[1,|\mathrm{D_i}|]\} \quad (3)$$

와 같이 정의한다. 이 때 $\mathrm{D_i}$는 i번째 은닉 영역이며,

b_{ij}는 D_i 영역의 j번째 4-문자 염기이고, $|D_i|$는 D_i의 염기 개수를 나타낸다. α는 은닉 영역의 최소 정수 부호의 개수이고, p는 예측 차수이다. 본 논문에서 최소 정수 부호의 개수 α를 10이상으로 설정하고, 예측 차수 p에 따라 은닉 영역을 선택한다.

3.1.3 가역 워터마크 은닉

n차 부호계수 히스토그램 정의역(domain) $Z=[0,2^{2n}-1]$은 M개 구간 $\{P_j\}_{j=1}^{M}$(section)으로 분할된다. 이 때 각 구간에서는 중간값 R_j을 기준으로 좌우 대칭이 되도록 구성되며, R_j은 쉬프팅의 기준값으로 사용된다. 구간의 길이는 홀수이며, 은닉 비트수 k에 의하여 결정된다. 구간 내 최대 쉬프팅 비트수가 k_{max}이고, 중간값이 $R_j=z$일 때, P_j는

$$P_j=\{z-2^{k_{max}}+1,\cdots,z-1,z,z+1,\cdots,z+2^{k_{max}}-1\},\ \text{for}\ j\in[1,M] \tag{4}$$

와 같이 $2\times2^{k_{max}}-1$개의 계수들로 구성되며, M은 구간 개수를 나타낸다. 이 때 잔여 구간은 $Z^c=Z-\sqcup_{j=1}^{M}P_j$으로 $2^{2n}-(2\times2^{k_{max}}-1)M$개 계수에 해당되며, 워터마크 은닉에 참여하지 않는다.

임의의 부호계수 x_i가 구간 P_j에 속할 경우, 구간 중간값 R_j과의 차이가 $d_i=x_i-R_j$, $x_i\in P_j$이다. 여기서 $|d_i|$의 범위에 따라 x_i 내에 은닉될 비트수 k_i가

$$\sum_{l=0}^{k_i-1}2^n<|d_i|\leqq\sum_{l=0}^{k_i}2^n,\ k_i\geqq1,\ \text{if}\ x_i\neq R_j \tag{5}$$
$$k_i=0,\ \text{if}\ x_i=R_j$$

와 같이 결정된 다음 x_i 내에 k_i비트 $\{w_l\}_{l=1}^{k_i}$가

$$x_i'=R_j+2^{k_i}d_i+\alpha(k_i)\ \ \text{where}$$
$$\alpha(k_i)=sgn(d_i)\sum_{l=1}^{k_i}2^{l-1}w_l, \tag{6}$$

와 같이 은닉된다.

임의의 히스토그램 정의역 구간 P_j (a section of histogram domain)는 Fig. 3에서와 같이 중간값 R_j 기준으로 왼쪽 서브구간 P_j^-과 오른쪽 서브구간 P_j^+으로 나뉘어진다. 구간 P_j의 오른쪽 서브구간 P_j^+ $(d>0)$(right subsection)에 해당되는 부호계수 x는 워터마크 비트에 의하여 오른쪽 구간 P_{j+1}(right section)의 왼쪽 서브구간 $P_{j+1}^-(d<0)$(left subsection)으로 이동된다. 이와 반대로 구간 P_j의 왼쪽 서브구간 $P_j^-(d<0)$에 해당되는 x는 워터마크 비트에 의하여 왼쪽 구간 P_{j-1}의 오른쪽 서브구간 P_{j-1}^+ $(d>0)$으로 이동된다. 이를 정리하면, Fig. 3(a)에서와 같이 구간 P_j의 오른쪽 서브구간의 부호계수와 오른쪽 인접 P_{j+1}의 왼쪽 서브구간의 부호계수과 서로 쉬프팅되며, 이와 반대로 구간 P_j의 왼쪽 서브구간의 부호계수와 왼쪽 인접 P_{j-1}의 오른쪽 서브구간의 부호계수가 서로 쉬프팅된다.

워터마크된 부호계수들 중에서 중간값 $x_i'=R_j$인 부호계수는 세 가지의 경우에 의하여 발생된다. 먼저 이전 부호계수가 중간값 $x_i=R_j$ $(k_i=0)$일 경우 쉬프팅에서 제외된다. 쉬프팅되지 않는 원래 부호계수 $x_i=R_j$이다. Fig. 3(a)에서와 같이 왼쪽 구간의 오른쪽 서브구간 P_{j-1}^+과 오른쪽 구간의 왼쪽 서브구간 P_{j+1}^- 내 계수들이 쉬프팅된 경우이다. 부호계수별 은닉 비트수에 의하여 쉬프팅되지 않은 경우와 쉬프팅된 경우를 구별할 수 있다. 추출 및 복원을 위하여 쉬프팅된 이전 구간 정보 $T=\{\tau\}$가

$$\tau=\begin{cases}0,\text{if}\ x'=R_j\ \text{and}\ x\in P_{j-1}^+\\1,\text{if}\ x'=R_j\ \text{and}\ x\in P_{j+1}^-\end{cases} \tag{7}$$

와 같이 필요하다. Fig. 3(b)에서와 같이 M개 구간 중에서 P_1의 오른쪽 서브구간 P_1^+에서부터 P_M의 왼쪽 서브구간 P_M^+내의 부호계수들이 쉬프팅이 된다. 나머지 경계 구간인 P_1^-와 P_M^+에 해당되는 부호계수들은 쉬프팅 비트수 $k=0$으로 할당된다.

구간 중간값 R_j에 해당되는 계수 $x_i=R_j$는 은닉 비트수 $k_i=0$으로 비트 은닉에서 제외된다. 여기서 쉬프팅된 부호계수 x_i'가 허위개시코돈 테이블 Z^c에 속하거나, x_i'와 이전 쉬프팅된 부호계수 x_i' 간의 허위개시코돈이 발생될 경우, 은닉 비트수 k_i가 0이 될 때까지 하나씩 감소하여 위의 과정을 반복한다. 이는 연속 부호계수 쌍 DE 방법과 동일하게 허위개시코돈을 방지한다. 이와 같은 방법으로 은닉 대상 영역의 모든 부호계수들에 대하여 부호계수별 은닉 비트수에 따라 다중비트가 은닉된 후, 워터마크된 비부호 영역 $I'(n)$가 얻어진다.

워터마크 추출 및 원본 서열 복원을 위한 부가정보로 부호계수별 은닉 비트수 $K=\{k_i\}$와 구간 기준값으로 쉬프팅된 구간 마커 $T=\{\tau\}$와 워터마크된 비부

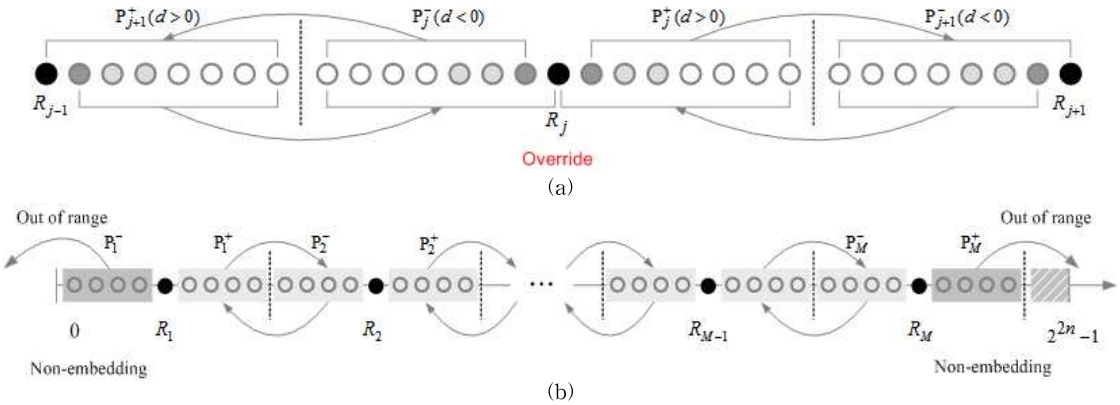

Fig. 3. (a) Code value shifting on a section P_j and neighborhood sections P_{j-1}, P_{j+1} and (b) code value shifting on each section and it's neighborhood sections on all sections.

호 영역 $\Gamma(n)$의 2비트 염기 이진수의 LSB 비트 B가 필요하다. 연속 부호계수 쌍 DE 방법과 동일하게 부가정보 (K,T,B)의 무손실 압축된 비트열 C을 생성한 다음, $\Gamma(n)$ 내 염기 이진수의 LSB 비트에 치환된다. 최종 부가정보 및 워터마크가 은닉된 비부호 영역 $\Gamma'(n)$을 가지는 DNA 서열 $D' = D'^{nc} + D^{c}$, , $D'^{nc} = \Gamma'(n) + \Gamma(n)$이 전송된다.

3.2 복호 및 복원 과정

제안한 복호 과정에서는 먼저 전송된 DNA 서열 D'의 비부호 영역 $\Gamma'(n)$로부터 압축 비트열의 부가정보 (K,T,B)를 얻는 다음, B의 염기 이진수 치환에 의하여 워터마크된 비부호 영역 $\Gamma'(n)$를 얻는다. $\Gamma'(n)$의 부호서열 X'로부터 부호계수별 쉬프팅 비트수 K와 쉬프팅 구간 마커 $T = \{\tau\}$에 의하여 워터마크와 원복 계수 복원이 수행된다.

부호서열 X'의 부호계수 x_i'가 주어졌을 때, 먼저 x_i'의 원래 구간의 중간값 R이 먼저 구하여야 한다. 즉, x_i'의 쉬프팅된 구간 P_j이 경계구간이 아니고 ($x_i' \in P_j$), 쉬프팅 비트수 k_i가 $k_i > 0$일 경우, x_i'의 이전 구간에 대한 중간값 R은

$$R = \begin{cases} R_{j-1}, \text{if } x_i' \in P_j^- \text{ or } \ddot{y} \ (x_i' = R_j \text{ and } \tau_i = 0) \\ R_{j+1}, \text{if } x_i' \in P_j^+ \text{ or } \ddot{y} \ (x_i' = R_j \text{ and } \tau_i = 1) \end{cases} \quad \text{if}$$

$$x_i' \in P_j \text{ and } k_i > 0 \tag{8}$$

와 같이 구해진다. 이 때 x_i'의 쉬프팅된 구간 P_j에 따라 은닉 전 구간의 중간값 R이 쉽게 구하여지나, x_i'가 쉬프팅된 영역 P_j의 중간값 R_j일 경우 ($x_i' = R_j$), 이전 구간 마커 τ_i에 의하여 R이 구해진다. x_i' 상에 워터마크 k_i비트 $\{w_l\}_{l=1}^{k_j}$와 원본 부호계수 x_i는 이전 구간의 중간값 R에 의하여

$$w_l = ((x_i' - R) \gg (l-1))\%2 \quad \text{for} \quad l = 1, \cdots, k_i \tag{9}$$

$$x_i = R + ((x_i'^{-} R) \gg k_i) \tag{10}$$

와 같이 구해진다.

3.3 워터마크 용량과 부가정보

부호차수 n과 구간 최대 쉬프팅 비트수 $k_{\max}$가 주어졌을 때, 은닉 영역 $\Gamma(n) = \{D_i\}_{i=1}^{|\Gamma(n)|}$내에 은닉되는 워터마크 비트수는 히스토그램 정의역 구간 P_j내에 중간값과의 차이 범위에 의하여 정의된 비트수와 각 구간에 부호계수가 속할 빈도수에 따라 결정되어진다.

부호계수 히스토그램 상에 z값을 가지는 빈도수를 $p(z)$이라 한다. 이 때 임의의 구간 P_j 상에 쉬프팅되는 비트수는 왼쪽 서브구간 P_j^-과 오른쪽 서브구간 P_j^+에 쉬프팅 비트수 $C(P_j^-)$와 $C(P_j^+)$의 합에 의하여 계산되어진다. $\Gamma(n) = \{D_i\}_{i=1}^{|\Gamma(n)|}$에 은닉된 총 워터마크 비트수는 전체 M개 구간 상에 경계구간 P_1^-와 P_M^+을 제외한 나머지 구간 상에 쉬프팅 비트수의 합이며, 염기 당 비트수 bpn $bpn_{HS}(n, k_{\max})$은

$$bpn_{HS}(n, k_{\max}) = \frac{1}{\sum_{i=1}^{|\Gamma(n)|} N_i} (C(P_1^+) \tag{11}$$

$$+\sum_{j=2}^{M-1}(C(\mathbf{P}_j^+)+\mathrm{C}(\mathbf{P}_j^-))+\mathrm{C}(\mathbf{P}_M^-)) \qquad \text{[bit/base]}$$

와 같이 정의된다. $|\Gamma(n)|$는 은닉 영역의 개수이고, N_i는 영역 D_i 내 부호계수의 개수로, $\sum_{i=1}^{|\Gamma(n)|}N_i$는 은닉 대상 영역의 총 염기의 개수이다. 워터마크 추출 및 복원을 위한 부가정보 $Extra_{\mathrm{HS}}(n,k_{\max})$는

$$Extra_{\mathrm{HS}}(n,k_{\max})=\mathrm{K}+\boldsymbol{T}+\mathrm{B} \qquad (12)$$

$$=\sum_{i=1}^{|\Gamma(n)|}(N_i(\lceil \log_2 k_{\max}\rceil+\sum_{j=1}^{M}p(x'=R_j))+|\mathrm{D}_i|) \qquad \text{[bit]}$$

와 같이 부호계수별 쉬프팅 비트수 K와 구간 기준값으로 쉬프팅된 구간 마커 $\boldsymbol{T}$와 워터마크된 비부호 영역 $\Gamma'(n)$의 2비트 염기 이진수의 LSB 비트 B의 합이다.

4. 실험결과

본 실험에서는 워터마크 용량, 부가데이터 용량, 염기변화율 및 허위개시코돈 발생 확률에 대하여 제안한 방법과 기존의 LS-PE 방법[9], Chen 방법[6], Huang 방법[7]과 비교 평가하였으며, NCBI Gen Bank에서 제공하는 DNA 서열을 사용했다. 각 DNA 서열의 타입, 접근 번호, 염기 개수, 비부호 DNA 영역 개수 정보는 Table 1에서와 같다.

제안한 방법의 용량은 부호차수 n과 최대 쉬프팅 비트수 $k_{\max}$에 의하여 결정되며, LS-PE 방법의 용량은 부호차수 n과 예측차수 p에 의하여 결정된다. Chen 방법과 Huang 방법은 부호에 필요한 염기 단위 $|w|$와 t에 의하여 용량이 결정되며, 이들은 $|w|=2$, $t=2$일 때 가장 높은 워터마크 용량을 가짐을 실험적으로 제시했다. 본 실험에서는 최대 워터마크 용량을 가지는 $k_{\max}$를 선택한 후, 각 방법들의 용량을 부호차수별로 비교 분석했다.

4.1 워터마크 bpn 대비 부가데이터 bpn

부호차수 n이 주어졌을 때, 제안한 방법의 구간별 부호계수 최대 쉬프팅될 수 있는 비트수는 $k_{\max}\leq 2n-2$이다. Fig. 4(a)는 $n\in[2,10]$일 때, $k_{\max}$가 $[2,\ 2n-2]$에 대한 제안한 방법의 워터마크 bpn를 보여준다. $n=2$, $k_{\max}=2$일 때 워터마크 bpn이 0.409bpn으로 가장 높게 나타났으며, n이 증가할수록 워터마크 bpn이 낮아진다. 실험결과로부터 부호차수 n에 따라 제안한 방법은 $k_{\max}=\lfloor 1/4(5n-2)\rfloor$일 때, 높은 워터마크 bpn를 가짐을 확인했다.

Fig. 4(b)의 워터마크 bpn 대비 부가데이터 bpn 결과를 살펴보면, 부호차수 n이 높을수록 부가정보량과 워터마크 bpn이 낮게 나타났다. $n=2$, $k_{\max}=2$일 때 제안한 방법은 워터마크 0.409bpn에 부가정보량

Table 1. Test DNA sequences

Type	Access No.	Total Bases	Noncoding DNA regions	Noncoding DNA bases
Archaea	AE017199	490,885	289	38,932
Bacterium	CP000108	2,572,079	1,770	301,761
Bacterium	CP000247	4,938,920	3,850	570,214
Bacterium	CP000672.1	1,887,192	1,444	466,266
Bacterium	AF012886.2	6,756	7	2,058
Bacterium	AE014075.1	5,231,428	3,767	631,026
Bacterium	CP000473.1	9,965,640	6,224	962,527
Eukaryota	nm_000520	2,437	2	847
Eukaryota	NC_001709.1	19,517	11	8,347
Eukaryota	NC_006033	1,195,132	533	393,739
Eukaryota	AL161582.2	198,669	302	137,622
Eukaryota	AL161595.2	198,151	215	126,917
Eukaryote	NC_006047	2,007,515	1,099	516,557
Moss	AP005672.1	122,890	99	51,916
Plant	NC_025652.1	141,255	68	91,971
Virus	AY653733.1	1,181,404	883	155,805

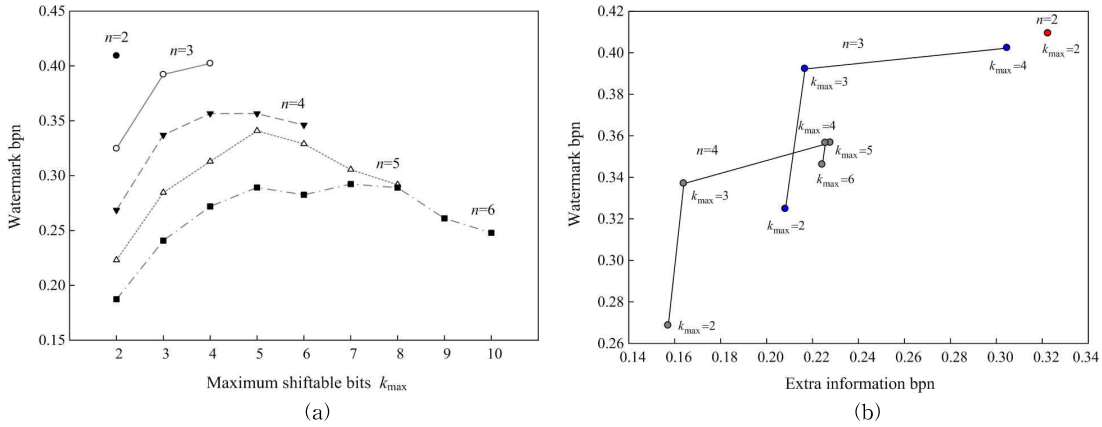

Fig. 4. (a) Watermark bpn and (b) extra information bpn vs watermark bpn for the code order n and the maximum shifting bits k_{max} of our method.

0.322bpn이 필요하다. 이는 LSB 치환 가능 비트량의 1/3배 정도에 해당되며, 약 3번 반복해서 은닉될 수 있는 부가정보량이다.

4.2 워터마크 bpn 대비 염기변화율

가역 DNA 워터마킹에서는 PSNR의 화질 척도와 유사한 염기변화율과 생물학적 기능 변경에 대한 평가가 필요하다. 본 실험에서는 워터마크에 의한 염기변화율을 비교 분석했다. 임의의 워터마크에 의하여 염기가 변경될 확률이 전체적으로 균등분포를 가진다고 가정할 때, 염기 변화율은 3/4=0.75에 가깝다.

Fig. 5(a)에서와 같이 제안한 방법의 염기변화율 결과를 살펴보면, 부호차수 n이 클수록 염기변화율이 높게 나타났다. 제안한 방법은 제일 높은 워터마크 bpn를 가지는 (n=2, k_{max}=2)일 때 염기변화율은 0.470으로, 낮은 워터마크 bpn를 가지는 (n=3, k_{max}=3), (n=4, k_{max}=3)의 염기변화율 0.501과 0.480보다 낮은 염기변화율을 가진다. Fig. 5(b)의 결과를 살펴보면, 제안한 방법은 0.46~0.68의 염기변화율을 가지면서, 0.28~0.41의 워터마크 bpn를 가진다. LS-PE 방법은 0.41~0.60의 염기변화율에 0.18~0.42의 워터마크 bpn를 가진다. 최대 높은 워터마크 bpn 기준으로 제안한 방법과 LS-PE 방법은 0.41bpn으로 유사한 워터마크 bpn를 가지나, 제안한 방법이 0.14 정도 낮은 염기변화율을 가진다.

4.3 허위개시코돈 방지

허위개시코돈 발생 확률 p_f는 비부호 워터마크 영역 D'^{nc}에서 임의의 연속 세 염기가 "ATG"가 될 확

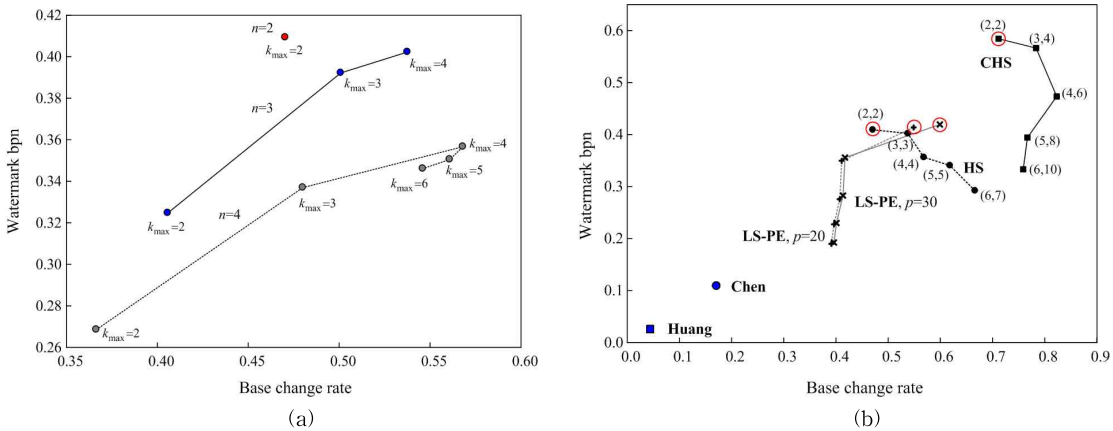

Fig. 5. Watermark bpn vs base change rate for (a) our method (n,k_{max}) and (b) conventional methods.

Table 2. Occurrence possibility of false start codon p_f

	Proposed method	LS-PE method [9]	Chen's method [6]	Huang's method [7]
p_f	0	0	9.11×10^{-5}	1.73×10^{-6}

률로 정의된다. 제안한 방법과 LS-PE 방법은 워터마크 은닉 과정과 부가 데이터의 LSB 치환 과정에서 허위개시코돈 방지를 위하여 비교 탐색 과정을 수행한다. 그러므로, 모든 실험 상에서 이 두 방법에서는 허위개시코돈이 발생되지 않았다. 허위개시코돈 발생에 대하여 고려하지 않는 Chen 방법과 Huang 방법은 실험 상에서 허위개시코돈이 발생되어, 생물학적 변이가 발생됨을 확인했다. 모든 테스트 DNA 서열의 1,000번 반복 수행하여 발생된 허위개시코돈 확률 p_f는 Table 2에서와 같이 나타내고 있다. Chen의 방법은 10^4 염기 당 하나의 허위개시코돈이 발생되며, Huang의 방법은 5.78×10^5 염기 당 하나의 허위개시코돈이 발생됨을 확인했다. 제안한 방법과 LS-PE 방법은 허위개시코돈이 발생되지 않았다.

5. 결 론

가역성 DNA 정보 은닉 기술은 워터마크된 DNA 서열 상에서 변이가 발생될 경우, 원본 DNA 서열로 복원으로 변이를 수정이 가능하다. 본 논문에서는 DNA 서열 중 아미노산으로 부호되지 않는 비부호 영역 내에 가역성 워터마크를 히스토그램 쉬프팅 기반으로 은닉하는 방법을 제안했다. 이 때 워터마크로 인하여 비부호 DNA 서열이 부호 DNA 서열로 변경되는 허위개시코돈이 발생될 수 있다. 제안한 방법에서는 DNA 정수열의 히스토그램 영역을 구간별로 선형적으로 구분한 다음, 경계 구간을 제외한 모든 구간의 히스토그램들을 인접 구간으로 워터마크 비트에 따라 서로 쉬프팅을 한다. 쉬프팅 과정에서 워터마크된 염기서열을 인접한 염기들 간의 비교 탐색을 통하여 허위개시코돈 생성을 방지한다. 기존 방법과의 비교 실험에 제안한 방법이 높은 워터마크 bpn를 가지는 것에 비하여 약 0.30 bpn 정도의 낮은 부가정보량이 필요함을 알 수 있었다. 특히 모든 실험 상에서 제안한 방법은 허위개시코돈이 전혀 발생되지 않음을 확인했다. 제안한 방법은 전염성 비디오 코덱 워터마킹[12] 상 가역성 워터마킹으로 적용 가능하다. 향후 본 연구진은 비가역성 부호 서열 워터마킹[13] 워터마킹을 가역성으로 변경하고자 한다.

REFERENCE

[1] C. Popovici, "Aspects of DNA Cryptography," *Mathematics and Computer Science Series*, Vol. 37, No. 3, pp. 147-151, 2010.

[2] D. Tulpan, C. Regoui, G. Durand, L. Belliveau, and S. Léger, "HyDEn: a Hybrid Steganocryptographic Approach for Data Encryption using Randomized Error-correcting DNA Codes," *BioMed Research International*, Article ID 634832, 2013.

[3] D. Heider and A. Barnekow, "DNA Watermarks-a Proof of Concept," *BMC Bioinformatics*, Vol. 9, No. 40, pp. 1-10, 2008.

[4] S.H. Lee, "DWT Based Coding DNA Watermarking for DNA Copyright Protection," *Information Sciences*, Vol. 273, pp. 263-286, 2014.

[5] S.H. Lee, "DNA Sequence Watermarking Based on Random Circular Angle," *Digital Signal Processing*, Vol. 25, pp. 173-189, 2014.

[6] T. Chen, "A Novel Biology-Based Reversible Data Hiding Fusion Scheme," *Frontiers in Algorithmics, Lecture Notes in Computer Science*, Vol. 4613, pp 84-95, 2007.

[7] Y.H. Huang, C.C. Chang, and C.Y. Wu, "A DNA-Based Data Hiding Technique with Low Modification Rates," *Multimedia Tools and Applications*, Vol. 70, Issue 3, pp. 1439-1451, 2014.

[8] S.H. Lee and K.R. Kwon, "Consecutive Difference Expansion Based Reversible DNA Watermarking," *The Institute of Electronics and Information Engineers*, Vol. 52, No. 7, pp. 51-62, 2015.

[9] S.H. Lee, S.G. Kwon, and K.R. Kwon, "Least

Square Prediction Error Expansion Based Reversible Watermarking for DNA Sequence," *The Institute of Electronics and Information Engineers*, Vol. 52, No. 11, pp. 66-78, 2015.

[10] R. Naskar and R.S. Chakraborty, "Histogram-Bin-Shifting-Based Reversible Watermarking for Colour Images," *Institution of Engineering Technology Image Processing,* Vol. 7, Issue 2, pp. 99-110, 2013.

[11] G. Coatrieux, W. Pan, N. Cuppens-Boulahia, F. Cuppens, and C. Roux, "Reversible Watermarking Based on Invariant Image Classification and Dynamic Histogram Shifting," *IEEE Transactions on Information Forensics and Security*, Vol. 8, Issue 1, pp. 111-120, 2013.

[12] B.J. Jang, S.H. Lee, and K.R. Kwon, "Reversible Watermarking based Video Contents Management and Control technique using Biological Organism Model," *Journal of Korea Multimedia Society*, Vol. 16, No. 7, pp. 841-851, 2013.

[13] S.H. Lee, S.G. Kwon, and K.R. Kwon, "DNA Watermarking Method based on Random Codon Circular Code," *Journal of Korea Multimedia Society*, Vol. 16, No. 3, pp. 318-329, 2013.

이 석 환

1999년 경북대학교 전자공학과 졸업(공학사)
2001년 경북대학교 전자공학과 졸업(공학석사)
2004년 경북대학교 전자공학과 졸업(공학박사)

2005년~현재 동명대학교 정보보호학과 부교수
관심분야 : 멀티미디어 보안, 영상신호처리, 바이오 보안

권 성 근

1996년 경북대학교 전자공학과 졸업(공학사)
1998년 경북대학교 전자공학과 졸업(공학석사)
2002년 경북대학교 전자공학과 졸업(공학박사)

2002년~2011년 삼성전자 무선통신사업부 연구원
2011년~현재 경일대학교 전자공학과 조교수
관심분야 : 멀티미디어 암호, 모바일 방송, 워터마킹

이 응 주

1992년 경북대학교 전자공학과 석사 졸업(공학사)
1996년 경북대학교 전자공학과 박사 졸업(공학박사)
2000년~2002년 ㈜디지털넷뱅크 대표이사

2005년~2006년 중국 대련공업대학교 초빙연구 교수
2005년~현재 한국멀티미디어학회 이사
1997년~현재 동명대학교 정보통신공학과 교수
관심분야 : 영상처리, 컴퓨터비전, 생체인식

권 기 룡

1986년 경북대학교 전자공학과 학사 졸업(공학사)
1990년 경북대학교 전자공학과 석사 졸업(공학석사)
1994년 경북대학교 전자공학과 박사 졸업(공학박사)

2000년~2001년 Univ. of Minnesota, Post-Doc.
1996년~2005년 부산외국어대학교 디지털정보공학부 부교수
2011년~2012년 Colorado State Univ., Visiting Scholar
2011년~2016년 IEEE R10 창원섹션 의장
2015년~2016년 한국멀티미디어학회 회장
관심분야 : 멀티미디어정보보호, 영상처리, 멀티미디어