Journal of The Institute of Electronics and Information Engineers Vol.53, NO.12, December 2016 https://doi.org/10.5573/ieie.2016.53.12.067
ISSN 2287-5026(Print) / ISSN 2288-159X(Online)

논문 2016-53-12-8

# 순환형 히스토그램 쉬프팅 기반 가역성 DNA 정보은닉 기법

## ( Reversible DNA Information Hiding based on Circular Histogram Shifting )

이 석 환*, 권 성 근**, 권 기 룡***

( Suk-Hwan Lee, Seong-Geun Kwon, and Ki-Ryong Kwon© )

요 약

DNA 컴퓨팅 기술로 DNA 정보를 매개물로 하는 DNA 저장, DNA 스테가노그라픽, 및 DNA 워터마킹에 대한 관심이 많아지고 있다. 생물학적 변이없이 외부 워터마크를 DNA 정보 내에 은닉에서는 원본 DNA 서열의 복원이 가능하고, 은닉과 복원이 반복적으로 이루어지며, 외부 워터마크에 의한 의도적인 변이 분석이 가능한 가역성 정보은닉 기술이 필요하다. 본 논문에서는 DNA 부호계수의 순환형 히스토그램 다중 쉬프팅 (Circular Histogram Shifting, CHS) 기반으로 생물학적 변이없이 허위개시코돈 방지, 원본 서열 길이 유지, 높은 워터마크 용량성, 블라인드 검출이 가능한 가역성 DNA 정보은닉 방법을 제안한다. 제안한 방법은 비부호 영역 DNA 염기서열을 부호계수로 변환한 다음, 높은 용량성을 위하여 순환형 히스토그램 다중 쉬프팅에 의하여 부호계수에 다중비트를 은닉한다. 마지막으로 다중비트 은닉 과정에서 은닉된 인접 염기서열 간의 비교탐색을 통하여 허위개시코돈 생성을 방지한다. 실험 결과로부터 제안한 방법이 기존 방법보다 0.11~0.50 bpn(bit per nucleotide base) 높은 워터마크 용량성을 가지고, 허위개시코돈이 발생되지 않음을 확인하였다.

Abstract

DNA computing technology makes the interests on DNA storage and DNA watermarking / steganography that use the DNA information as a newly medium. DNA watermarking that embeds the external watermark into DNA information without the biological mutation needs the reversibility for the perfect recovery of host DNA, the continuous embedding and detecting processing, and the mutation analysis by the watermark. In this paper, we propose a reversible DNA watermarking based on circular histogram shifting of DNA code values with the prevention of false start codon, the preservation of DNA sequence length, and the high watermark capacity, and the blind detection. Our method has the following features. The first is to encode nucleotide bases of 4-character variable to integer code values by code order. It makes the signal processing of DNA sequence easy. The second is to embed the multiple bits of watermark into -order coded value by using circular histogram shifting. The third is to check the possibility of false start codon in the inter or intra code values. Experimental results verified the our method has higher watermark capacity 0.11~0.50 bpn than conventional methods and also the false start codon has not happened in our method.

**Keywords** : Reversible DNA watermarking, Histogram shifting, DNA security, Bio-security

## Ⅰ. 서 론

최근 DNA 컴퓨팅 기술로 인하여 DNA를 매개물로 하는 DNA 저장[1~2] 및 비밀 통신, 암호화 및 저작권 보호를 위한 DNA 워터마킹[3~7]에 대하여 많은 연구가 이루어지고 있다. DNA 저장 및 워터마킹에서는 원본 DNA 서열의 손실없이 복구할 수 있는 가역성 기술이 매우 필요하다. 대부분의 DNA 워터마킹 방법[3~7]들은

* 정회원, 동명대학교 정보보호학과 (Dept. of Information Security, Tongmyong Univ.)
** 정회원, 경일대학교, 전자공학과 (Dept. of Electronic Engineering, KyungIl Univ.)
*** 정회원, 부경대학교, IT융합응용공학과 (Dept. of IT Convergence and Application Eng., Pukyong National Univ.)
© Corresponding Author(E-mail: krkwon@pknu.ac.kr)
※ 본 연구는 정부(교육과학기술부)의 재원으로 한국연구재단의 기초연구사업 지원 (NRF-2011-0023118, NRF-2014R1A1A4A01006663, NRF-2016R1D1A3B03931003)과 2013년도 부산광역시 Brain Busan (BB21) 사업 지원을 받아 수행된 것임.
Received ; August 9, 2016 Revised ; September 7, 2016 Accepted ; November 16, 2016

원본 복구 기능이 없는 비가역에 해당된다.

가역 DNA 정보은닉은 DNA 저장, DNA 서열의 위변조 방지 뿐만 아니라 외부 워터마크에 의한 생물학적 변이를 방지하는 한편, 은닉과 복원의 반복적 과정을 통하여 워터마크에 의한 생물학적 변이 과정 분석이 가능하다. 그러나 가역 DNA 정보은닉은 일반 멀티미디어 데이터와는 달리 네 개의 문자로 구성된 염기 서열의 낮은 신호량으로 인하여 비가역 DNA 정보은닉에 비하여 연구가 많이 진행되고 있지 않다.

기존의 가역 DNA 정보은닉 방법들을 살펴보면, Chen 등[8]은 비부호 DNA 서열의 네 문자열을 십진수로 변환한 후, 무손실 압축 및 DE(Difference Expansion) 기반 방법을 적용하였다. Huang 등[9]은 낮은 염기 변화율을 위하여 히스토그램 기반 방법을 적용하였으나, 낮은 용량을 가진다. Liu 등[10]은 상보쌍 염기 치환 기반 데이터 은닉으로 워터마크 추출 및 복원시 참조되는 원본 서열이 필요하다. 최근 Lee 등은 가역성 영상 정보은닉에서 사용되는 부호계수 쌍의 차분 확장(DE, Difference expansion)[11] 및 최소자승 오차(LS-PE, Least square prediction error) 확장 기반[12]으로 데이터를 은닉하였다. 일반적으로 인접 화소 간의 유사성이 많은 영상데이터와는 달리 염기서열에서는 인접 부호계수 간의 유사성이 크지 않으므로, 선형 예측 기반 차분 확장 방법보다 히스토그램 쉬프팅 방법이 보다 많은 용량을 가진다. 이상의 기존 방법들은 원본 서열 길이를 유지하나, 허위개시코돈 방지를 고려하지 않고, 비블라인드이거나, 낮은 용량을 가진다.

본 논문에서는 허위개시코돈 방지 및 높은 용량성을 위한 순환형 히스토그램 다중 쉬프팅 기반 비부호 DNA 서열의 가역 정보은닉 방법들을 제시한다. 제안한 방법에서는 4문자-염기서열을 연속적인 $n$개 염기 단위(또는 $n$ 부호 차수)의 정수형 부호계수열로 변환한다. 이 때, 인접 구간 및 최소 및 최대 영역의 경계 구간 간에 쉬프팅이 되도록 부호계수열을 순환형 히스토그램(Circular Histogram)으로 할당한다. 그리고 워터마크 메시지의 다중비트 단위로 히스토그램 쉬프팅을 수행한 후, 워터마크된 염기서열 내에 인접 염기서열 간의 비교 탐색을 통하여 개시코돈 생성을 방지한다. 실험 결과로부터 제안한 방법의 워터마크 용량이 LS-PE 방법[12]보다 0.120~0.194 bpn 정도 높고, Chen 방법[8]과 Huang 방법[9]보다 0.5 bpn 정도 높음을 확인하였다. 또한 기존 Chen 방법[8]과 Huang 방법[9]은 허이개시코돈이 발생되었으나, 제안한 방법과 LS-PE 방법[12]에서는 허위개시코돈이 발생되지 않음을 확인하였다.

본 논문의 구성은 다음과 같다. 먼저 2장에서는 가역 DNA 정보은닉 이론과 기존 방법에 대하여 살펴본다. 3장에서는 제안한 염기서열 부호화, 허위개시코돈 방지에 대하여 살펴본 후, 순환형 히스토그램 다중 쉬프팅 방법에 대하여 자세히 살펴본다. 4장에서는 제안한 방법과 기존 방법과의 비교 실험 분석 한 후, 마지막 5장에서 본 논문의 결론을 맺는다.

## Ⅱ. 관련 연구

### 1. 가역성 DNA 정보은닉 이론

DNA 서열은 멀티미디어 데이터의 화질과 같은 속성은 없으나, 고려해야 될 주요 특징이 있다.

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. 따라서 염기 서열에 대한 신호 처리가 제한적이므로, 이로 인한 은닉 데이터의 용량이 낮다. 그러나 연속된 염기 서열을 조합할 경우 다중 비트의 정보를 가지게 된다. 예를 들어, 4개의 연속된 염기 서열은 44=256 레벨을 가지게 되므로, 8비트 영상 화소값과 같은 레벨을 가진다.

두 번째로는 은닉 과정에 의하여 비부호 DNA의 일부 염기 서열이 부호 DNA 서열의 시작 코돈(Methionine, ‘ATG’)으로 변경될 가능성이 있다. 따라서 워터마크 은닉 과정에서 시작 코돈으로 변경될 염기 서열들을 예측하여 제외하여야 한다.

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

### 2. 기존 연구

본 장에서는 기존의 가역성 DNA 정보은닉 방법에 대하여 살펴보기로 한다. Chen 등[8]은 “ATCG”의 염기 심볼들을 2비트 이진으로 변환한 다음, 염기 이진 서열을 임의의 비트 단위의 워드로 십진수로 변환한다. 그리고 십진수 서열의 워드 쌍을 확장 집합 S1, 변경 집

합 S2, 비변경 집합 S3으로 분류한 후 각 워드 쌍의 위치맵을 생성한다. 압축된 위치맵(map), S2에 속한 워드 쌍의 원본 LSB들(LSB(S2)), 및 비밀 메시지들을 압축열의 마지막 단에 추가한다. 그리고 S1에 속한 워드 쌍의 차이에 비밀메시지 비트가 은닉된다. 이 방법은 Thodi 등[13]의 가역성 영상 워터마킹 방법을 적용한 것으로, 은닉 데이터 용량이 매우 낮은 평균 0.09-0.13 bpn를 가진다.

Huang 등[9]은 Chen의 방법과 같이 이진 염기서열을 비트 단위로 십진수로 변환한 다음, 십진수 서열의 히스토그램을 구한다. 이 때 h를 가장 높은 빈도수의 값, L1을 가장 낮은 빈도수의 값, L2를 두 번째 낮은 빈도수의 값이라 한다. 임의의 십진수 $p_j$가 L1이면, L2로 변경하고, 위치맵을 1로 놓는다. 이와 반대로 $p_j$가 L2이면, $p_j$를 변경하지 않고, 위치맵을 0으로 놓는다. $p_j$가 h일 때, 은닉 비트가 0이면, $p_j$는 변경하지 않고, 은닉 비트가 1이면, $p_j$는 L1으로 변경한다. 이 방법은 염기 변경율은 낮으나, bpn이 매우 낮고, Chen의 방법과 같이 허위개시코돈의 발생된다.

Lee 등은 DNA 서열의 부호계수열이 예측이 어려운 랜덤한 균등 분포를 가지므로, 인접 부호계수 쌍의 차분 확장 (DE)[11]과 최소자승 예측오차 확장 (LS-PE)[12] 기반으로 다중 비트를 은닉하는 방법을 제안하였다. DE 방법은 인접 부호계수 쌍의 차분 확장에 은닉할 최대 비트를 은닉하고, LS-PE 방법은 현재 부호계수에 대한 최소자승 예측 오차를 구한 다음, 예측오차 확장 조건에 따라 결정된 비트수만큼 예측오차를 확장한다. LS-PE 방법은 DE 방법보다 약 1.1배 정도 높은 0.4 bpn를 가지며, 인접 염기서열 간의 비교탐색을 통하여 허위개시코돈 생성을 방지한다. 그러나 워터마크 추출에 필요한 부가데이터 정보가 1.11 bpn 정도 필요하며, 인접 계수 쌍에 대한 차이 확장으로 고용량 데이터 은닉에는 적합하지 않다. 따라서 낮은 부가데이터와 높은 워터마크 용량을 가지는 가역성 DNA 정보은닉 방법이 필요하다.

## Ⅲ. 제안한 DNA 가역 정보은닉

본 논문에서는 순환형 히스토그램 쉬프팅을 이용한 비부호 DNA 가역 정보은닉 기법을 제안한다. DNA 가역 정보은닉은 그림 1에서와 같이, 워터마크가 은닉되기에 적절한 길이를 가지는 인트론(Intron) 성분으로 구

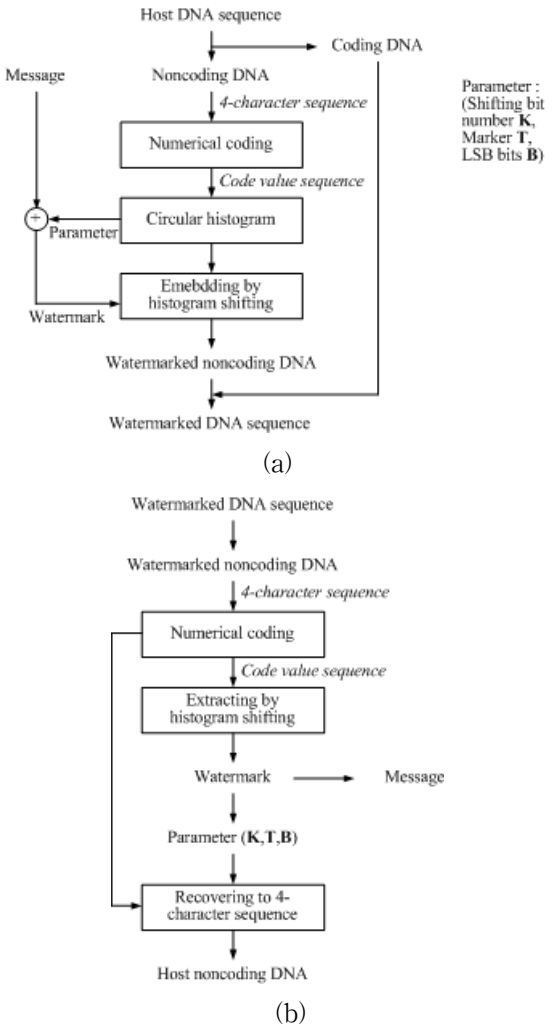



그림 1. 제안한 (a) 가역성 메시지 은닉 과정 및 (b) 메시지 추출 및 원본 DNA 서열 복원 과정

Fig. 1. Proposed processes for (a) reversible message embedding and (b) message extracting and host DNA sequence recovering.

성된 비부호 DNA 영역에 워터마크가 은닉되며, 은닉된 영역들은 추출 및 복원 과정에 의하여 워터마크가 추출되고, 원본 영역으로 복원된다.

### 1. 가역성 메시지 은닉

#### 1) 4-문자 부호화

4-문자 염기 서열에 대한 신호 처리 용이성을 위하여 다중비트 부호 과정이 필수적이다. 제안한 방법에서는 LS-PE 방법[12]에서 제시된 다중비트 부호 과정에 의하여 DNA 염기 서열을 부호한다.

일반적으로 뉴클레오티드 염기는 b=(A, T, C, G) 4-문자로 표현되며, 이를 4개의 십진수 또는 2비트의 이진수로 $b=(0,1,2,3)_{10}=(00,01,10,11)_2$와 같이 표현된다. 여기서 $n$개 염기들로 구성된 염기 블록

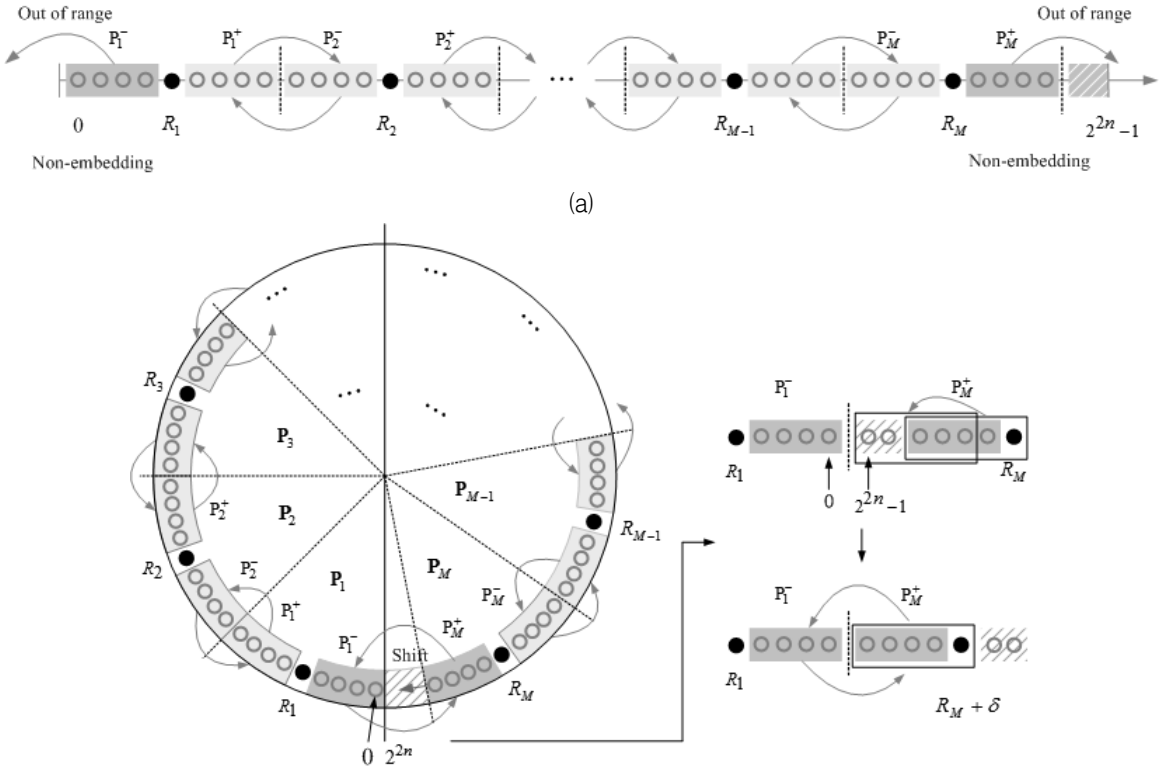



그림 2. (a) 전체 구간 상에서 각 구간과 좌우 인접 구간 간의 부호계수 쉬프팅 및 (b) 구간별 순환형 히스토그램 다중 쉬프팅

Fig. 2. (a) Shifting of histogram values on all sections with boundary sections and (b) Circular histogram multiple shifting on sections.

$\mathbf{x}=(\mathrm{b}_1,\mathrm{b}_2,\cdots,\mathrm{b}_n)$ 단위로 $2n$비트의 부호계수 $x$로

$$x=f(\mathbf{x})=\sum_{k=1}^{n}(b_k\cdot 2^{2(n-k)}) \tag{1}$$

와 같이 부호된다. 부호계수 $x$로부터 염기 블록의 염기들은 $f^{-1}(x)=\mathbf{x}$, $b_k=(x\gg 2(n-k))\%4$와 같이 쉽게 복원되어진다. 본 논문에서는 염기 블록의 염기 개수 $n$를 부호차수로 부르기로 한다. 은닉 영역 $\mathbf{D}_k$ 내 염기들은 부호차수 $n$에 의하여 부호계수 $\mathbf{X}_k$로 부호된다; $\mathbf{X}_k=\{x_i|i\in[1,N_k]\}$. $N_k=\lfloor|\mathbf{D}_k|/n\rfloor$. 이 때 부호계수 개수 $N_k$은 부호차수 $n$에 의하여 결정된다.

2) 가역성 메시지 은닉

비부호 영역의 부호계수들은 영상의 화소값과는 달리 화질에 대한 조건이 없으므로, 최대치와 최소치 간의 쉬프팅이 가능하다. 제안한 방법에서는 그림 2(a)에서와 같이 $n$차 부호계수 히스토그램 정의역(domain) $\mathbf{Z}=[0,2^{2n}-1]$ 을 $M$개 구간 $\{\mathbf{P}_j\}_{j=1}^{M}$으로 분할한다. 이 때 각 구간에서는 중간값 $R_j$을 기준으로 좌우 대칭이 되도록 구성되며, $R_j$은 쉬프팅의 기준값으로 사용된다. 따라서 구간의 길이는 홀수이며, 은닉 비트수 $k$에 의하여 결정된다. 구간 내 최대 쉬프팅 비트수가 $k_{\max}$이고, 중간값이 $R_j=z$일 때, 임의의 구간 $\mathbf{P}_j$는 $\mathbf{P}_j=\{z-2^{k_{\max}}+1,\cdots,z-1,z,z+1,\cdots,z+2^{k_{\max}}-1\}$와 같다.

워터마크 용량을 높이기 위하여, 경계 구간에 해당되는 $\mathbf{P}_1$의 왼쪽 서브구간 $\mathbf{P}_1^-(d<0)$과 $\mathbf{P}_M$의 오른쪽 서브구간 $\mathbf{P}_M^+(d>0)$에서도 은닉이 가능하도록 그림 2(b)에서와 같이 히스토그램 구간 쉬프팅을 순환형으로 변경한다. 먼저, 서브구간 $\mathbf{P}_M^+$는 잔여 구간으로 쉬프팅되어, $\mathbf{P}_M$의 두 서브구간들이 분리된다. 잔여 구간의 부호계수 개수가 $\delta=2^{2n}-(2\times 2^{k_{\max}}-1)M$일 때, 구간 $\mathbf{P}_M$ 영역은

$$\mathbf{P}_M=\mathbf{P}_M^-+\mathbf{P}_M^+ \tag{2}$$

와 같이 왼쪽 서브구간 $\mathbf{P}_M^-$과 오른쪽 서브구간 $\mathbf{P}_M^+$으로

$$\mathbf{P}_M^- = \left\{z - 2^{k_{\max}} + 1, \cdots, z-1, z\right\}, \quad (3)$$

$$\mathbf{P}_M^+ = \left\{z+\delta, z+\delta+1, \cdots, z+\delta+2^{k_{\max}} - 1 (= 2^{2n}-1)\right\} \quad (4)$$

나누어진다. 여기서 $\mathbf{P}_M^-$와 $\mathbf{P}_M^+$의 기준값은 각각 $R_M^- = z$, $R_M^+ = z+\delta$와 같다. 여기서 잔여 구간은 히스토그램 내의 영역이나, 정수개의 구간 분할에 의하여 남겨진 구간이다.

임의의 부호계수 $x_i$ 상에서 $x_i$가 속한 구간 $\mathbf{P}_j$의 중간값 $R$

$$R = \begin{cases} R_j, & \text{if } x_i \in P_j \quad \text{for } j = 1,2,\cdots,M-1 \\ R_M^-, & \text{if } x_i \in P_M^- \text{ for } j = M \\ R_M^+, & \text{if } x_i \in P_M^+ \text{ for } j = M \end{cases} \quad (5)$$

에 의하여 $k_i$비트 $\{w_n\}_{n=1}^{k_i}$가

$$x_i^{'} = (R + 2^{k_i} d_i + \alpha(k_i)) \,\%\, 2^{2n} \quad (6)$$

where $d_i = x_i - R$, $\alpha(k_i) = sgn(d_i)\sum_{l=1}^{k_i} 2^{l-1} w_l$

와 같이 은닉된다. 이 때 $\mathbf{P}_M^-$과 $\mathbf{P}_M^+$ 사이의 잔여 계수 $[R_M^- + 1, R_M^+ - 1]$와 각 구간의 중간값에 해당되는 부호계수들은 쉬프팅 비트수가 0이다.

인접 구간의 중간값으로 쉬프팅된 계수 $x_i^{'}$에 대한 이전 구간에 대한 정보 $\mathbf{T}$는

$$t = \begin{cases} 0, \text{if } (x' = R_j \text{ and } x \in \mathbf{P}_{j-1}) \text{ or} \\ \qquad (x' = R_M^+ \text{ and } x \in \mathbf{P}_1) \\ 1, \text{if } (x' = R_j \text{ and } x \in \mathbf{P}_{j+1}) \text{ or} \\ \qquad (x' = R_1 \text{ and } x \in \mathbf{P}_M^+) \end{cases} \quad (7)$$

와 같이 결정된다.

이와 같은 방법으로 부호서열 $\mathbf{X}$의 모든 부호계수에 워터마크가 은닉되며, 워터마크된 비부호 영역 $\Gamma'(n)$가 얻어진다. 워터마크 복호 및 원본 부호계수의 복원에 필요한 부가정보는 비순환형 방법과 동일한 부호계수별 쉬프팅 비트수 $\mathbf{K}$와 쉬프팅 구간 마커 $\mathbf{T}$와 2비트 엠기 이진수의 LSB 비트 $\mathbf{B}$이다. 압축된 부가정보의 LSB 치환은 위의 두 방법과 동일하게 적용되며, 치환된 영역 $\Gamma''(n)$에 의하여 최종 워터마크된 DNA 서열 $\mathbf{D}'$을 전송한다.

## 2. 메시지 추출 및 DNA 복원 과정

전송된 DNA 서열의 치환 영역 $\Gamma''(n)$으로부터 역치환에 의하여 워터마크된 영역 $\Gamma'(n)$이 얻어진 다음, $\Gamma'(n)$ 내의 부호서열 $\mathbf{X}'$로부터 $(\mathbf{K},\mathbf{T})$에 의하여 워터마크가 복호되며, 원본 부호서열이 복원된다.

부호서열 $\mathbf{X}'$ 내에 $k_i > 0$인 부호계수 $x_i^{'}$가 주어졌을 때, 경계구간과 비경계구간에 따라 $x_i^{'}$의 이전 구간의 중간값 $R$을 구하여야 한다. $x_i^{'}$이 비경계구간에 속할 경우

$$R = \begin{cases} R_{j-1}, \text{if } x_i^{'} \in P_j^- \text{ or } (x_i^{'} = R_j \text{ and } t_i = 0) \\ R_{j+1}, \text{if } x_i^{'} \in P_j^+ \text{ or } (x_i^{'} = R_j \text{ and } t_i = 1) \end{cases} \quad (8)$$

와 경계구간에 속할 경우

$$R = \begin{cases} R_M^+, \text{if } 0 \leqq x_i^{'} < R_1 \text{ or } (x_i^{'} = R_1 \text{ and } t_i = 0) \\ R_1, \quad \text{if } R_M^+ < x_i^{'} \leqq 2^{2n-1} \text{ or} \\ \qquad (x_i^{'} = R_M^+ \text{ and } b_i = 1) \end{cases} \quad (9)$$

에 따라 중간값 $R$이 구하여진다. 그리고 $R$에 의하여 $k_i$비트 $\{w_l\}_{l=1}^{k_i}$와

$$w_l = (((x_i^{'} - R) \% 2^{2n}) \gg (l-1)) \% 2 \quad (10)$$

$$\text{for } l = 1, \cdots, k_i$$

와 같이 구하여지며, 원본 부호계수 $x_i$는

$$x_i = R + ((x_i^{'} - R) \% 2^{2n} \gg k_i) \quad (11)$$

와 같이 복원된다.

## 3. 워터마크 용량 및 부가정보

순환형 히스토그램 쉬프팅 방법에서는 부호계수 히스토그램 정의역 범위에 잔여 구간을 제외한 모든 구간에서 워터마크가 은닉된다. 따라서 부호차수 $n$과 구간 최대 쉬프팅 비트수 $k_{\max}$가 주어졌을 때, 은닉 영역 $\Gamma(n)$ 내에 워터마크 비트수는 각 구간의 왼쪽 서브구간 $\mathbf{P}_j^-(d<0)$과 오른쪽 서브구간 $\mathbf{P}_j^+(d>0)$상에 쉬프팅 비트수의 합으로 이에 대한 bpn $bpn(n,k_{\max})$은

$$bpn(n,k_{\max}) = \frac{1}{\sum_{i=1}^{|\Gamma(n)|} N_i} \sum_{j=1}^{M} (C(\mathbf{P}_j^+) + C(\mathbf{P}_j^-)) \quad (12)$$

와 같다. 워터마크 추출 및 복원을 위한 부가정보 $Extra(n,k_{\max})$는 부호계수별 쉬프팅 비트수 $\mathbf{K}$와 구간 기준값으로 쉬프팅된 구간 마커 $\mathbf{T}$와 워터마크된 비

부호 영역 $\Gamma'(n)$의 2비트 염기 이진수의 LSB 비트 B이다.

$$Extra(n,k_{\max}) = \mathrm{K} + \mathrm{T} + \mathrm{B} \quad (13)$$

여기서 히스토그램 정의역 구간 내의 최대 쉬프팅 비트수가 $k_{\max}$일 때, 은닉 비트수는 $\lceil \log_2 k_{ma} \rceil$ 비트로 표현되므로, 전체 부호계수에 대한 쉬프팅 비트수 K는 $\mathrm{K} = \lceil \log_2 k_{\max} \rceil \sum_{i=1}^{|\Gamma(n)|} N_i$와 같다. 쉬프팅 구간 마커 T는 인접 구간의 중간값으로 쉬프팅된 부호계수 $x' = R_j$가 왼쪽 또는 오른쪽 구간에서 쉬프팅된 것인지 판별하는 이진 정보로 $\mathrm{T} = \sum_{i=1}^{|\Gamma(n)|} N_i \times \sum_{j=1}^{M} p(x' = R_j)$ 비트로 표현된다. 그리고 B는 $\Gamma'(n)$ 내 모든 영역의 염기 개수와 동일한 $\mathrm{B} = \sum_{i=1}^{|\Gamma(n)|} |\mathrm{D}_i|$ 비트이다.

## Ⅳ. 실험 결과

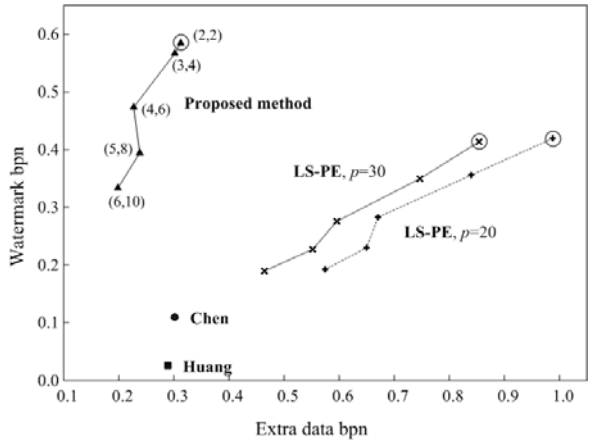
영상 가역 워터마킹의 성능 평가에서는 용량(bpp; bit per pixel)에 대한 PSNR이 사용되나, DNA 가역 정보은닉에서는 PSNR의 화질 척도 대신 생물학적 기능 변경이 사용된다. 제안한 방법에서는 비부호 영역에 허위 개시코돈 발생되지 않도록 가역 워터마크를 은닉하므로 생물학적 기능 변경이 없다. 따라서 본 실험에서는 제안한 순환형 히스토그램 쉬프팅 방법과, Chen 방법[8], Huang 방법[9], LS-PE 방법[12]의 워터마크 용량 $bpn$, 압축 부가정보량 $Extra$, 염기 변화율, 및 허위개시코돈 발생 가능성에 대하여 비교 분석하였다. 실험에서 사용된 DNA 서열은 NCBI GenBank에서 제공된 것이며, 이들의 타입, Access No., 염기 개수, 비부호 DNA 영역 개수 정보는 표 1에 나타내었다. 비부호 영역들은 다양한 염기 개수를 가지며, 매우 작은 염기 수로 구성된 영역들은 제안한 영역 선택 과정에 의하여 은닉 대상에서 제외된다.

### 1. 워터마크 용량

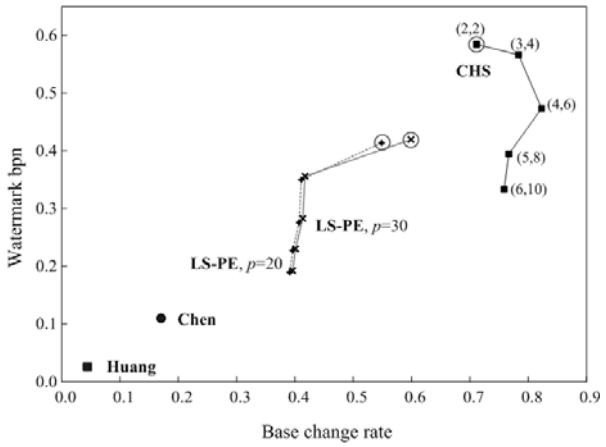
본 실험에서는 제안한 방법의 최대 쉬프팅 비트수 $k_{\max}$를 $k_{\max} = 2n-2$ $(n \in [2,6])$로 설정한 다음, 기존 방법의 워터마크 bpn를 비교 평가하였다. 워터마크 bpn과 부가데이터 bpn의 실험 결과는 표 1과 그림 3에 나타내었다.

표 1. 테스트 DNA 서열에 대한 워터마크 bpn 및 부가데이터 bpn
Table 1. Watermark bpn and extra information bpn for test DNA sequences.

| Type | Access No. | Total Bases | Noncoding 염기 개수 | 제안 방법 $(n,k_{\max})$=(2,2) | | LS-PE 방법[12] $(n,p)$=(2,30) | | Chen 방법[8] $(\lvert w\rvert$=2) | | Huang 방법[9] $(t$=2) | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | 워터마크 bpn | 부가데이터 bpn | 워터마크 bpn | 부가데이터 bpn | 워터마크 bpn | 부가데이터 bpn | 워터마크 bpn | 부가데이터 bpn |
| Archaea | AE017199 | 490,885 | 38,932 | **0.581** | 0.316 | 0.425 | 0.930 | 0.098 | 0.232 | 0.034 | 0.202 |
| Bacterium | CP000108 | 2,572,079 | 301,761 | **0.577** | 0.312 | 0.414 | 1.159 | 0.114 | 0.207 | 0.015 | 0.161 |
| Bacterium | CP000247 | 4,938,920 | 570,214 | **0.582** | 0.313 | 0.412 | 1.194 | 0.096 | 0.218 | 0.021 | 0.208 |
| Bacterium | CP000672.1 | 1,887,192 | 466,266 | **0.585** | 0.313 | 0.401 | 0.670 | 0.098 | 0.222 | 0.044 | 0.170 |
| Bacterium | AF012886.2 | 6,756 | 2,058 | **0.604** | 0.343 | 0.429 | 5.300 | 0.111 | 0.213 | 0.022 | 0.242 |
| Bacterium | AE014075.1 | 5,231,428 | 631,026 | **0.584** | 0.312 | 0.404 | 1.023 | 0.115 | 0.216 | 0.044 | 0.207 |
| Bacterium | CP000473.1 | 9,965,640 | 962,527 | **0.573** | 0.311 | 0.405 | 1.048 | 0.115 | 0.196 | 0.037 | 0.215 |
| Eukaryota | nm_000520 | 2,437 | 847 | **0.528** | 0.389 | 0.409 | 10.601 | 0.099 | 0.217 | 0.028 | 0.175 |
| Eukaryota | NC_006033 | 1,195,132 | 393,739 | **0.591** | 0.315 | 0.414 | 0.672 | 0.117 | 0.207 | 0.025 | 0.213 |
| Eukaryota | AL161582.2 | 198,669 | 137,622 | **0.590** | 0.314 | 0.422 | 0.611 | 0.115 | 0.210 | 0.046 | 0.223 |
| Eukaryota | AL161595.2 | 198,151 | 126,917 | **0.592** | 0.315 | 0.419 | 0.559 | 0.097 | 0.235 | 0.003 | 0.198 |
| Eukaryote | NC_006047 | 2,007,515 | 516,557 | **0.600** | 0.315 | 0.411 | 0.793 | 0.115 | 0.217 | 0.016 | 0.221 |
| Moss | AP005672.1 | 122,890 | 51,916 | **0.589** | 0.314 | 0.455 | 0.662 | 0.109 | 0.227 | 0.012 | 0.204 |
| Plant | NC_025652.1 | 141,255 | 91,971 | **0.579** | 0.316 | 0.398 | 0.470 | 0.116 | 0.213 | 0.037 | 0.183 |
| Virus | AY653733.1 | 1,181,404 | 155,805 | **0.603** | 0.316 | 0.463 | 1.131 | 0.101 | 0.237 | 0.015 | 0.227 |
| 평균 | | | | **0.584** | 0.321 | 0.419 | 1.788 | 0.108 | 0.218 | 0.027 | 0.203 |
| 용량효율=부가데이터/워터마크 | | | | 0.550 | | 4.271 | | 2.021 | | 7.643 | |




(a)




(b)

그림 3. 제안한 방법 $(n,k_{\max})$과 기존 Chen, Huang 방법, LS-PE 방법 ($p$=20,30)에 대한 (a) 워터마크 bpn vs 부가정보량 (b) 워터마크 bpn vs 염기변화율

Fig. 3. (a) Watermark bpn vs extra information bpn and (b) watermark bpn vs base change rate of proposed method $(n,k_{\max})$, Chen's method, Huang's method, and LS-PE method ($p$=20,30).

그림 3(a)의 워터마크 bpn를 살펴보면, 제안한 방법은 $(n,k_{\max})=(2,2)$일 때 0.584 bpn로 제일 높으며, LS-PE 방법은 $p$=30일 때 0.419 bpn, $p$=20일 때 0.413 bpn,으로 높게 나타났다. 기존의 Chen 방법은 0.108 bpn, Huang 방법은 0.027 bpn으로 제안한 방법에 비하여 매우 낮게 나타났다.

워터마크 bpn 대비 부가정보량을 살펴보면, LS-PE 방법은 제안한 방법에 비하여 0.265~0.542 bpn 정도 높은 부가정보량이 필요하다. Chen과 Huang의 방법은 낮은 워터마크 bpn에 비하여 0.30 정도의 부가정보량 필요하다. 따라서 제안한 방법이 다른 방법에 비하여 가장 높은 워터마크 bpn를 가지는 것에 비하여 0.3 정도의 낮은 부가정보량이 필요함을 알 수 있었다.

## 2. 부가정보량

그림 3(a)의 부가데이터 bpn를 살펴보면, Chen과 Huang 방법이 각각 0.218 bpn, 0.203 bpn 정도이고, 제안한 방법은 각각 0.331 bpn, 0.321 bpn 정도이다. LS-PE 방법은 1.788 bpn으로 매우 높은 부가데이터가 필요하다. 기존 방법들은 낮은 워터마크 용량으로 낮은 부가데이터가 필요하다. 그러나 워터마크 bpn 대비 부가데이터 bpn의 용량효율을 살펴보면, 제안한 방법이 0.550으로 워터마크 대비 가장 낮은 부가데이터가 필요함을 볼 수 있다.

## 3. 염기변화율

염기 변화율은 워터마크에 의하여 변경된 염기의 비율을 나타낸다. 임의의 워터마크에 의하여 염기가 변경될 확률이 전체적으로 균등분포를 가진다고 가정할 때, 염기 변화율은 3/4=0.75에 가깝다.

그림 3은 염기변화율에 대한 워터마크 bpn를 나타낸다. 제안한 방법은 0.7~0.8의 염기변화율을 가지면서, 0.32~0.58의 워터마크 bpn를 가진다. LS-PE 방법은 0.4~0.6의 염기변화율에 0.18~0.42의 다소 낮은 워터마크 bpn를 가진다. 즉, 제안한 방법은 기존 LS-PE 방법에 비하여 0.14~0.16 높은 워터마크 bpn를 가지나, 0.28~0.38 정도 높은 염기변화율을 가진다. 그러나 비부호 DNA 서열은/ 정크 DNA으로 가정할 때, 염기 변화율이 높더라도 부호 DNA 서열에 영향을 미치지 않는다. 또한 제안한 방법들은 부호 DNA 서열의 시작 코돈으로 변경되지 않으므로, 아미노산 서열에 전혀 영향을 미치지 않는다. 즉, 데이터 은닉된 서열 $\Gamma'(n)$을 가지는 비부호 DNA 서열은 $\mathbf{D}'^{nc}=\Gamma'(n)+\Gamma^{C}(n)$와 같으며, 은닉된 DNA 서열은 $\mathbf{D}'=\mathbf{D}'^{nc}+\mathbf{D}^{c}$와 같다. 따라서 아미노산 변화율은 0이다.

만약 비부호 DNA 서열에서 발현 인자 조절과 같은 주요한 염기 서열이 알려졌을 경우, 허위개시코돈 방지와 같은 방법으로 이들 서열이 변경되지 않도록 한다.

## 4. 허위 개시코돈 방지

제안한 방법은 워터마크 은닉 과정과 부가 데이터의 LSB 치환 과정에서 허위 개시코돈 방지를 위하여 비교 탐색 과정을 수행한다. 따라서 모든 실험 상에서 제안한 방법이 허위 개시코돈이 발생되지 않음을 확인하였다. 그러나 기존의 Chen 방법과 Huang 방법은 허위 개시코돈 발생에 대하여 고려하지 않으며, 실험의 워터마크 은닉 과정에서 허위 개시코돈이 발생되었다.

비부호 워터마크 영역 $\mathbf{D}'^{nc}$에서 임의의 연속 세 염기가 "ATG"가 될 확률

$$p_f = P(b_i b_{i+1} b_{i+2} = "ATG" | \mathbf{D}'^{nc}) \quad (14)$$

을 허위 개시코돈 발생 확률이라고 한다. 모든 테스트 DNA 서열의 1,000번 반복 수행하여 발생된 허위 개시코돈 확률 $p_f$는 표 2에서와 같이 나타내고 있다. Chen의 방법은 $10^4$ 염기 당 하나의 허위 개시코돈이 발생되며, Huang의 방법은 $5.78\times10^5$ 염기 당 하나의 허위 개시코돈이 발생됨을 확인하였다. 제안한 DE, HS, 및 CHS 방법 모두 허위 개시코돈이 발생되지 않았다.

표 2. 허위개시코돈 발생 확률 $p_f$
Table 2. Occurrence probability $p_f$ of false start codon.

| 구분 | 제안 방법 | LS-PE | Chen | Huang |
|---|---|---|---|---|
| $p_f$ | 0 | 0 | $9.11\times10^{-5}$ | $1.73\times10^{-6}$ |

## V. 결 론

본 논문에서는 DNA 저장, 스테가노그라픽, 및 워터마킹 등의 바이오 정보보호를 위하여 순환형 히스토그램 다중 쉬프팅를 이용한 가역성 DNA 정보은닉 방법을 제안하였다. DNA 정보는 외부 워터마크에 의하여 생물학적 변이를 유발할 수 있으므로, 정보은닉된 DNA 정보로부터 원본 DNA가 오류없이 복원이 되어야 한다. 제안한 방법은 4-문자 염기서열을 부호차수에 의한 정수형 부호계수로 변환한 다음, 순환형 히스토그램 다중 쉬프팅을 이용하여 부호계수별 다중비트를 은닉한다. 일반적인 가역 영상 워터마킹은 영상 화질에 의하여 다중 쉬프팅이 어렵다. 그러나 DNA 염기서열은 영상 화질과 같은 가시성 기준이 없는 반면, 생물학적 변이 유지 및 허위개시코돈 방지 등의 제한 조건 하에서 DNA 부호계수의 변경이 자유롭다. 실험 결과로부터 제안한 방법은 기존의 LS-PE 방법, Chen 방법, Huang 방법보다 약 0.11~0.50 bpn 정도의 워터마크 bpn이 높았으며, 허위개시코돈이 발생되지 않음을 확인하였다.

## REFERENCES

[1] G. M. Church, Y. Gao, S. Kosuri, "Next-generation digital information storage in DNA," *Science*, Vol. 337, No. 6102, pp. 1628, Sep. 2012.
[2] N. Goldman, P. Bertone, S. Chen, C. Dessimoz, E. M. LeProust, B. Sipos, and E. Birney, "Towards practical high-capacity, low-maintenance information storage in synthesized DNA," *Nature*, Vol. 494, pp. 77-80, Feb. 2013.
[3] D. Heider and A. Barnekow, "DNA-based watermarks using the DNA-Crypt algorithm," *BMC Bioinformatics*, Vol. 8, No. 176, May 2007.
[4] M. Liss, D. Daubert, K. Brunner, K. Kliche, U. Hammes, A. Leiherer, and R. Wagner, "Embedding permanent watermarks in synthetic genes," *PLOS* ONE, Vol. 7, Issue 8, e42465, Aug. 2012.
[5] S. H. Lee, "DWT based coding DNA watermarking for DNA copyright protection," *Information Sciences*, Vol. 273, pp. 263-286, July, 2014.
[6] S. H. Lee, "DNA sequence watermarking based on random circular angle," *Digital Signal Processing*, Vol. 25, pp. 173-189, Feb. 2014.
[7] S.-H. Lee, K.-R. Kwon, and S.-G. Kwon, "A Robust DNA Watermarking in Lifting Based 1D DWT Domain," *Journal of the Institute of Electronics and Information Engineers*, Vol. 49, No. 10, pp. 91-101, Oct. 2015.
[8] 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.
[9] Y.-H. Huang, C.-C. Chang, and C.-Y. Wu, "A DNA-based data hiding technique with low modification rates," *Multimedia Tools and Applications*, Volume 70, Issue 3, pp 1439-1451, June 2014.
[10] G. Liu, H. Liu, and A. Kadir, "Hiding message into DNA sequence through DNA coding and chaotic map," *Medical & Biological Engineering & Computing*, vol. 52, issue 9, pp. 741-747, Sep. 2014.
[11] S.-H. Lee and K.-R. Kwon, "Consecutive difference expansion based reversible dna watermarking," *Journal of the Institute of Electronics and Information Engineers*, Vol. 52, No. 7, pp. 51-62, July 2015.
[12] S.-H. Lee, S.-G. Kwon, and K.-R. Kwon, "Least Square Prediction Error Expansion Based Reversible Watermarking for DNA Sequence," *Journal of the Institute of Electronics and Information Engineers*, Vol. 52, No. 11, pp. 66-78, Nov. 2015.
[13] D. M. Thodi et al. "Expansion embedding techniques for reversible watermarking," *IEEE Trans. on Image Processing*, Vol. 16, No. 3, March 2007.

## 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 : 영상신호처리, 컨텐츠보안, 3D그래픽스>

**권 성 근**(정회원)
1996년 경북대학교 전자공학과 학사 졸업.
1998년 경북대학교 전자공학과 석사 졸업.
2002년 경북대학교 전자공학과 박사 졸업.
2002년～2011년 삼성전자 무선사업부 책임연구원
2011년～현재 경일대학교 전자공학과 부교수
<주관심분야 : 멀티미디어 암호, 모바일 방송, 워터마킹>

**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 컴퓨터전자공학과 부교수
2006년～현재 부경대학교 IT융합응용공학과 교수
<주관심분야 : 통신, 컴퓨터, 신호처리, 반도체>