Journal of The Institute of Electronics and Information Engineers Vol.52, NO.11, November 2015 http://dx.doi.org/10.5573/ieie.2015.52.11.066
ISSN 2287-5026(Print) / ISSN 2288-159X(Online)

논문 2015-52-11-8

최소자승 예측오차 확장 기반 가역성 DNA 워터마킹

(Least Square Prediction Error Expansion Based Reversible Watermarking for DNA Sequence)

이 석 환*, 권 성 근**, 권 기 룡***

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

요 약

바이오컴퓨팅 기술이 발전함에 따라 DNA 정보를 매개물로 한 DNA 워터마킹에 대한 연구가 이루어지고 있다. 그러나 DNA 정보는 일반 멀티미디어 데이터와는 달리 생물학적으로 중요한 정보이므로, 원본 DNA가 복원이 되는 가역성 DNA 워터마킹 기술이 필요하다. 본 논문에서는 최소자승 (Least square) 예측오차 확장 (prediction error expansion) 기반으로 비부호 DNA 서열의 가역성 워터마킹 기법을 제안한다. 제안한 방법에서는 비부호 영역의 4-문자 염기서열들을 인접한 개 염기에 의한 정수형 부호계수로 변환한다. 그리고 현재 부호계수에 대한 최소자승 예측 오차를 구한 다음, 예측오차 확장 조건에 따라 결정된 비트수만큼 예측오차를 확장한다. 이때 은닉된 인접 염기서열 간의 비교탐색을 통하여 허위개시코돈 생성을 방지한다. 실험 결과로부터 제안한 예측오차 확장 방법이 기존 방법과 평균 예측오차 확장 방법보다 높은 워터마크 용량을 가지며, 생물학적 변이 및 허위개시코돈이 발생되지 않음을 확인하였다.

Abstract

With the development of bio computing technology, DNA watermarking to do as a medium of DNA information has been researched in the latest time. However, DNA information is very important in biologic function unlikely multimedia data. Therefore, the reversible DNA watermarking is required for the host DNA information to be perfectively recovered. This paper presents a reversible DNA watermarking using least square based prediction error expansion for noncodng DNA sequence. Our method has three features. The first thing is to encode the character string (A,T,C,G) of nucleotide bases in noncoding region to integer code values by grouping n nucleotide bases. The second thing is to expand the prediction error based on least square (LS) as much as the expandable bits. The last thing is to prevent the false start codon using the comparison searching of adjacent watermarked code values. Experimental results verified that our method has more high embedding capacity than conventional methods and mean prediction method and also makes the prevention of false start codon and the preservation of amino acids.

Keywords : 가역 DNA 워터마킹, 바이오 보안, DNA 보안, 최소자승 예측오차

* 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)과 2013년도 부산광역시 Brain Busan (BB21) 사업 지원을 받아 수행된 것임.
Received ; June 23 2015 Revised ; October 2, 2015 Accepted ; November 3, 2015

Ⅰ. 서 론

최근 DNA 저장[1~2] 뿐만 아니라, DNA를 매개물로 하는 비밀 통신과 암호화를 위한 DNA 스테가노그라픽[3]-[5], DNA 서열의 저작권 보호를 위한 DNA 워터마킹[6]-[11]에 대하여 많은 연구가 이루어지고 있다. DNA 저장/스테가노그라픽/워터마킹에서 필요한 공통 주제로 생물학적 기능 변경없이 DNA 서열 내에 외부 정보를 어떤 목적으로 어떻게 은닉하는 것이다. 이에 따라 본 연구진은 부호 영역의 코돈 서열에 대하여 64개의 랜덤 원형 각도 변환[9] 및 코돈 서열 DWT 기반의 비가역 워터마킹 기법[10~11]을 제안하였다. 그러나 DNA 워터마킹에서는 원본 DNA 서열의 손실없이 복구할 수 있는 가역 워터마킹 기술이 매우 필요하다. 가역 DNA 워터마킹은 일반 멀티미디어 데이터와는 달리 네 개의 문자로 구성된 염기 서열의 낮은 신호량으로 비가역 DNA 워터마킹에 비하여 연구가 많이 진행되고 있지 않다.

DNA 서열은 부호 영역(Coding DNA)과 비부호 영역(Noncoding DNA)으로 나누어진다. 부호 DNA은 단백질로 부호화되는 컴포넌트로, 워터마크 은닉시 단백질로의 부호가 변경되지 않아야 한다. 비부호 DNA은 단백질로 부호하지 않은 컴포넌트로, 대부분 Junk DNA로 알려져 있다. 가역 DNA 워터마킹은 원본 DNA 서열의 손실없이 복구가 가능한 것으로, 부호 영역보다 비부호 영역에 대하여 적용이 가능하다.

기존의 가역 DNA 워터마킹 방법[12~15]들을 살펴보면, Chen 등[12]은 비부호 DNA 서열의 네 문자열을 십진수로 변환한 후, 대표적인 가역 영상 워터마킹 방법인 무손실 압축 및 DE(Difference Expansion) 기반 방법을 적용하였다. Huang 등[13]은 낮은 염기 변화율을 위하여 히스토그램 기반 방법을 적용하였으나, 낮은 용량을 가진다. Liu 등[14]은 상보쌍 염기 치환 기반 데이터 은닉으로 워터마크 추출 및 복원시 참조되는 원본 서열이 필요하다. 위의 방법들은 원본 서열 길이를 유지하나, 허위개시코돈 방지를 고려하지 않고, 비블라인드이거나, 낮은 용량을 가진다. 기존 방법의 낮은 용량 문제점을 해결하기 위하여 본 연구진은 연속적인 부호계수의 차이를 다중비트 확장 방법[15]을 제안하였다. 이 방법은 랜덤계수 부호계수 예측의 어려움으로 인접 계수 간의 차이를 이용한 것으로, 위의 두 방법보다 높은 워터마크 용량을 가지며, 허위개시코돈이 발생되지 않으나 부호계수 예측에 대한 필요성을 제기하였다.

가역 DNA 워터마킹과 가역 영상 워터마킹[16~19]와의 주요 차이점은 다음과 같다. 1) 인접 화소 간의 유사성이 많은 영상과는 달리 염기서열에서는 인접 부호계수 간의 유사성이 크지 않다. 따라서 가역성 영상 워터마킹에서 많이 사용되는 인접 간의 유사성을 이용한 예측오차 확장[18~19]은 염기서열에 적합하지 않다. 2) 염기서열의 부호계수는 생물학적 기능 변경 또는 허위개시코돈 발생이 되지 않는 한 동적 범위 내에 이동이 가능하다. 즉, 가역 영상 워터마킹에서는 화질적인 측면이 매우 주요한 고려사항이나, 가역 DNA 워터마킹에서는 (A,T,C,G)의 4-문자 내에 이동이 자유로우나, 아미노산 유지와 허위개시코돈 방지 등의 제한 조건을 가진다.

본 논문에서는 허위개시코돈 방지, 블라인드, 높은 용량성을 위한 최소자승 (least square, LS) 예측오차 확장 기반의 비부호 DNA 서열의 가역 정보은닉 방법들을 제안한다. 제안한 방법은 부호계수 변환, 예측오차 확장, 허위개시코돈 방지의 세 단계로 구성된다. 먼저 부호계수 변환 과정에서는 워터마킹 신호처리 용이를 위하여 4-문자 염기서열을 연속적인 n개 염기 단위(또는 n 부호차수)의 정수형의 부호계수열로 변환한다. 이때 n 부호차수에 따라 2^{2n}비트의 부호계수열가 생성된다. 예측오차 확장 과정에서는 예측차수 p의 최소자승 예측기를 이용하여 부호계수별 예측계수 오차를 구한 다음, 최대 확장이 가능한 비트수만큼 예측오차를 확장한다. 예측오차 확장에 의하여 워터마크된 부호계수와 4-문자 염기서열이 차례로 구하여진다. 마지막 과정에서는 워터마크된 염기서열 내 (인트라 모드) 및 인접 염기서열 간 (인터 모드)의 비교 탐색을 통하여 시작코돈 생성을 방지한다.

실험에서는 제안한 LS 기반 예측오차 확장 방법과 기존의 Chen 방법[12], Huang 방법[13], 연속 부호계수 차이 방법[15] 및 평균 예측오차 확장 방법과의 워터마크 용량 bpn (bit per nucleotide base)과 허위개시코돈 발생 확률을 비교하였다. 실험 결과로부터 제안한 방법이 부호차수와 예측차수가 (n,p)=(2,2)일 때, 평균 예측오차 방법보다 약 1.08배, Chen의 방법보다 약 3.88배, Huang 방법보다 약 15.5배 정도 워터마크 용량이 많음을 확인하였다. 또한 기존 방법들은 1.73×10^{-6}~9.11×10^{-5} 확률의 허위개시코돈이 발생되어 비부호 영역이 부호 영역으로 인식되어 허위의 아미노산이 발생되

었다. 그러나 제안한 방법은 허위개시코돈이 전혀 발생되지 않음을 확인하였다.

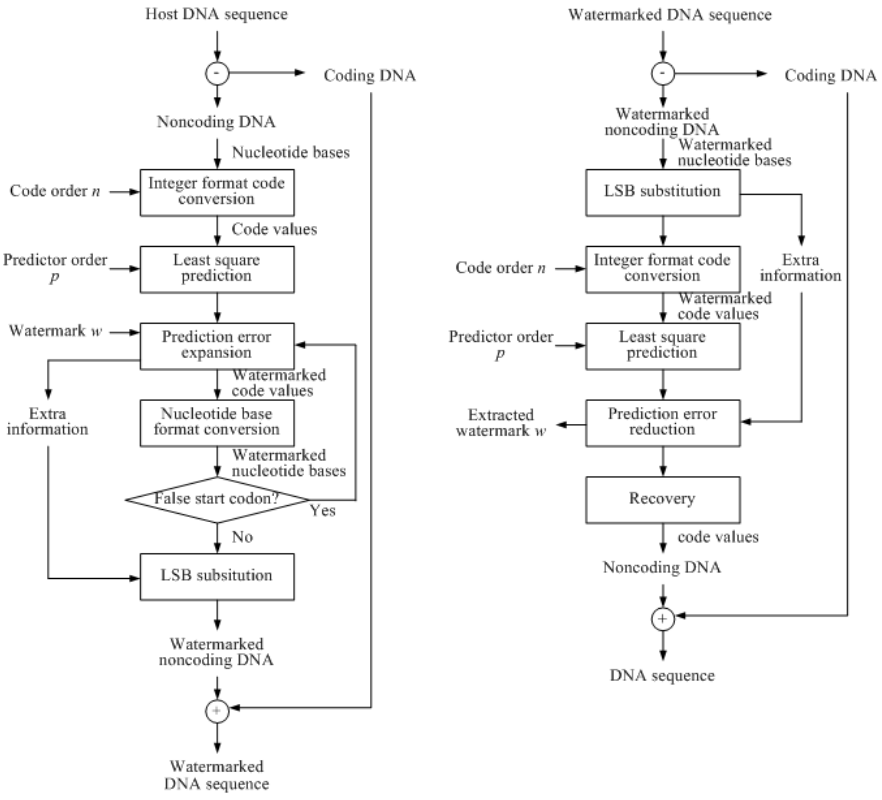
본 논문의 구성을 살펴보면, Ⅱ장에서는 기존 가역 DNA 정보은닉 방법에 대하여 살펴본 다음 Ⅲ장에서는 제안한 염기서열 부호화, 허위개시코돈 방지 및 LS 기반 예측오차 확장 방법에 대하여 자세히 살펴본다. Ⅳ장에서는 제안한 네 방법과 기존 방법과의 비교 실험 분석 한 후, 마지막 Ⅴ장에서는 본 논문의 결론을 맺는다.

Ⅱ. 기존 가역 DNA 워터마킹

비부호 DNA은 멀티미디어 데이터의 화질과 같은 속성은 없으나, 허위개시코돈 방지 및 블라인드 검출 및 서열 길이 보존 등을 고려하여야 한다. 허위개시코돈은 은닉 과정에 의하여 비부호 DNA의 일부 염기 서열이 부호 DNA 서열의 개시 코돈(Methionine, 'ATG')으로 변경될 가능성이 있다. 따라서 워터마크 은닉 과정에서 개시코돈으로 변경될 염기서열들을 제외하여야 한다.

가역성 DNA 워터마킹 방법으로 Chen 등[12]은 이들은 염기 이진 서열을 $|w|$ 비트 단위의 십진수로 변환 후 산술 부호화에 의하여 압축한 다음, 이진 비밀 메시지를 압축열에 추가한다. 이 방법은 $|w|$가 2비트일 때 bpn이 제일 높으며, 실험적으로 평균 0.75~0.81bpn를 가진다. 그러나 이 방법은 압축열에 비밀 메시지가 추가되므로, 압축열 길이가 증가된다. 또한 비부호 DNA 특성을 고려하지 않으므로, 허위개시 코돈이 발생되며, 은닉 데이터 용량이 매우 낮은 단점을 가진다.

Huang 등[13]은 염기서열을 $2t$비트 단위의 십진수로 변환한 다음, 십진수 서열의 히스토그램을 구한다. 이때 h를 가장 높은 빈도수의 값, L1을 가장 낮은 빈도수의 값, L2를 두 번째 낮은 빈도수의 값이라 한다. 임의의 십진수 p_j가 L1이면, p_j를 L2로 변경하고, 위치맵을 1로 놓는다. p_j가 L2이면, p_j를 변경하지 않고, 위치맵



(a) (b)

그림 1. 제안한 가역 DNA 워터마킹; (a) 워터마크 삽입 과정, (b) 워터마크 추출 및 DNA 서열 복원 과정
Fig. 1. Proposed reversible DNA watermarking;
(a) watermark embedding process and (b) detecting and recovering process.

을 0으로 놓는다. p_j가 h일 때, 은닉 비트가 0이면, p_j는 변경하지 않고, 은닉 비트가 1이면, p_j는 L1으로 변경한다. 복원과 추출은 위치맵, h, L1, L2 값에 의하여 수행된다. 이 방법은 염기 변경율과 워터마크 bpn이 낮고, Chen의 방법과 같이 허위개시 코돈의 발생된다.

Liu 등[14]은 비밀 메시지를 Chebyshev chaotic 맵에 의하여 암호화한 다음, PWLCM (Piecewise linear chaotic map)에 의하여 은닉 위치를 선정한다. 그런 다음, 상보 쌍 규칙에 의하여 염기 당 2비트를 은닉한다. 이 방법들은 상보 쌍 치환 방법에 의하여 데이터를 은닉하는 것으로, 추출 및 복원을 위하여 참조 또는 원본 DNA 서열이 필요한 비블라인드 방법이다.

Lee 등[15]은 DNA 서열의 부호계수열이 예측이 어려운 랜덤한 균등 분포를 가지므로, 인접 부호계수 쌍의 차분 확장에 다중 비트를 은닉하는 방법을 제안하였다. 이 방법은 기존 방법의 비블라인드, 허위개시코돈 발생 등의 문제를 해결하였다. 그러나 부가데이터 정보가 필요하며, 인접 계수 쌍에 대한 차이 확장으로 고용량 데이터 은닉에는 적합하지 않다. 따라서 본 논문에서는 고용량의 데이터 은닉을 위하여 LS 기반의 균등분포 부호계수 예측을 이용한다.

Ⅲ. 제안한 LS 예측 기반 가역 DNA 워터마킹

제안한 워터마크 삽입 과정과 추출 및 DNA 서열 복원 과정은 그림 1에서와 같다. DNA 서열은 비부호 영역들 $\mathbf{D}^{nc}$과 부호 영역들 $\mathbf{D}^{c}$로 구성되며, 비부호 영역들 $\mathbf{D}^{nc}$ 중에 워터마크가 은닉되기에 적절한 길이를 가지는 은닉 영역 Γ이 선택된다.

1. 부호차수에 의한 염기 정수 부호화

4-문자 염기 서열에 대한 워터마킹 신호 처리 용이성을 위하여 다중비트 부호 과정이 필수적이다. 일반적으로 뉴클레오티드 염기는 ATCG의 4-문자로 표현되며, 이를 4개의 십진수 또는 2비트의 이진수로 표현된다; $b=(0,1,2,3)_{10}=(00,01,10,11)_2 \leftarrow$ b=(A,T,C,G). 신호 처리 용이성을 위하여 2비트 계수보다 그림 2에서와 같이 2비트 이상의 다중비트로 표현된 계수로 확장하여야 한다. 제안한 방법에서는 n개 염기들로 구성된 염기 블록 $\mathbf{x}$ 단위로 $2n$비트의 부호계수 x로

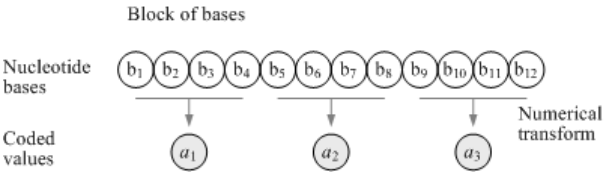


그림 2. 부호차수 n=4일 때 n차 염기 블록에 대한 $2n$ 비트의 계수 표현 ($a_k \in [0, 2^{2\times 4}-1]$)

Fig. 2. Numerical representation of $2n$ bits coded values for n bases in code order n=4.

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

where $\mathbf{x}=(b_1, b_2, \cdots, b_n)$, $x \in [0, 2^{2n}-1]$

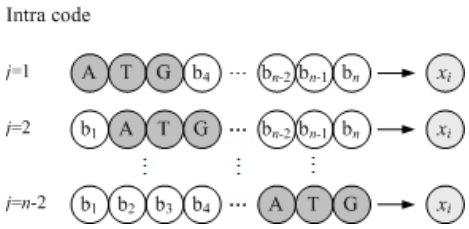
와 같이 부호한다. 부호계수 x로부터 염기들은

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

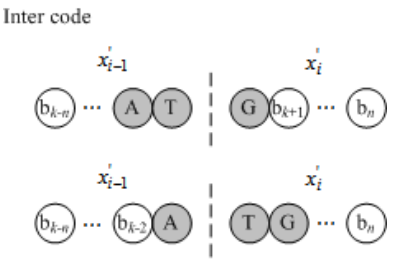
와 같이 쉽게 복원되어진다. 은닉 영역 $\mathbf{D}_k$ 내 염기들은 부호차수 n에 의하여 부호계수 $\mathbf{X}_k$로 부호된다; $\mathbf{X}_k = \{x_i | i \in [1, N_k]\}$. $N_k = \lfloor |\mathbf{D}_k|/n \rfloor$.

2. 허위개시코돈 방지

워터마크 은닉 과정에서 일부 염기 서열은 부호 영역



(a)



(b)

그림 3. (a) 부호계수 내 및 (b) 부호계수 간에 허위개시 코돈

Fig. 3. False start codon (a) in coded value and (b) between coded values.

의 개시코돈인 'ATG'으로 변경될 수 있다. 이와 같은 허위개시코돈은 비부호 영역의 일부가 부호 영역으로 변경하며, 생물학적 기능의 변경을 유발시킨다. 허위개시코돈은 다음과 같이 부호계수 내 또는 부호계수 간에 발생될 수 있다.

1) 부호계수 내 (Intra code) : 부호차수 $n>2$일 경우, 그림 3(a)에서와 같이 부호계수 정의역 내에 $n-2(n>2)$개의 허위개시코돈들이 발생될 수 있다. 염기 블록 내에 임의의 위치 $j\in[1,n-2]$에서 발생된 허위개시코돈을 포함하는 부호계수는 $2^{2(n-3)}$개이므로 $n-2$개의 위치에서 발생된 허위개시코돈을 포함하는 부호계수는 총 $(n-2)\times 2^{2(n-3)}$개다. 제안한 방법에서는 염기 부호화에서 허위개시코돈을 포함하는 부호계수 테이블 $\mathbf{Z}^c$을 미리 생성한 다음, 워터마크된 부호계수 x'가 $\mathbf{Z}^c$에 포함되지 않도록 은닉 과정을 수행한다.

2) 부호계수 간 (Inter codes) : 이전 워터마크된 부호계수의 염기 블록 $\mathbf{x}_{i-1}^{'}$과 현재 부호계수의 염기 블록 $\mathbf{x}_i^{'}$ 간에 허위개시코돈이 발생될 수 있다. 그림 3(b)에서와 같이 $(\mathbf{x}_{i-1}^{'}\mathbf{x}_i^{'})$ 가 (…A, TG…) 또는 (…AT, G…)일 때 중간 위치에 허위개시코돈이 발생된다. 제안한 방법에서는 이전 워터마크된 부호계수 $x_{i-1}^{'}$가 주어질 때, 현재 부호계수 x_i에 대한 은닉 비트수를 조절한다.

3. 다중비트 은닉위한 부호계수 오차 확장 조건

인접 화소 간의 상관관계가 높은 영상 데이터에서는 화소(또는 계수) 쌍에 k비트(k>1) 은닉시, 예측오차 (d)를 2^k배 확장하여야 하므로, 두 화소 간의 화질 열화가 발생된다. 즉, 화질 열화에 따라 1비트 이상의 데이터 은닉이 어렵다. 이와는 달리 화질에 대한 조건이 없는 DNA 부호계수는 허위개시코돈 계수를 제외한 유효 범위 내에 이동이 자유롭다. 따라서 부호계수 쌍에 대한 예측오차 (d)는 확장 조건에 따라 k 비트 은닉을 위하여 2^k배 확장이 가능하며, 최대 $2n-1$비트 은닉이 가능하다; $k_{\max}=2n-1$.

워터마크의 k비트 $\{w_j\}_1^k$와 예측계수 $\hat{x}$가 주어졌을 때, k 비트 은닉된 부호계수 x'는

$$x'=\hat{x}+2^k d+sgn(d)\sum_{j=1}^{k}2^{j-1}w_j \tag{3}$$

와 같이 2^k배 확장된 예측오차 $d=x-\hat{x}$에 의하여 구하여진다. 은닉된 부호계수 x'와 비트 수 k가 주어졌을 때, 워터마크 추출 및 복원은

$$w_j=((x'-\hat{x})\gg(j-1))\%2 \text{ for } j=1,\cdots,k \tag{4}$$

$$x=\hat{x}+d=\hat{x}+(x'-\hat{x})\gg k \tag{5}$$

와 같이 쉽게 구하여진다.

은닉된 부호계수 x'는 $0\leq x'\leq 2^{2n}-1$이어야 하므로, 2^k배 확장위한 예측오차 d의 확장 조건은

$$2^{-k}(-\hat{x}-\alpha(k))\leq d\leq 2^{-k}(2^{2n}-1-\hat{x}-\alpha(k)) \tag{6}$$

$$x\in[\max(0,\lceil\hat{x}+2^{-k}(-\hat{x}-\alpha(k))\rceil),\ \min(2^{2n}-1,\lfloor\hat{x}+2^{-k}(2^{2n}-1-\hat{x}-\alpha(k)\rfloor))] \tag{7}$$

where $\alpha(k)=sgn(d)\sum_{j=1}^{k}2^{j-1}w_j$

와 같다. 이와 같은 확장 조건은 워터마크 k비트 $\{w_j\}_1^k$와 예측계수 $\hat{x}$에 의하여 결정되며, 확장 조건에 따라 부호계수 x에 은닉될 비트수가 결정된다.

그림 4는 부호차수 n=4 ($x,\hat{x}\in[1,2^8-1]$)이고 워터마크 비트가 전부 1일 때 $w=\{1\}$, 예측계수 $\hat{x}$별 부호

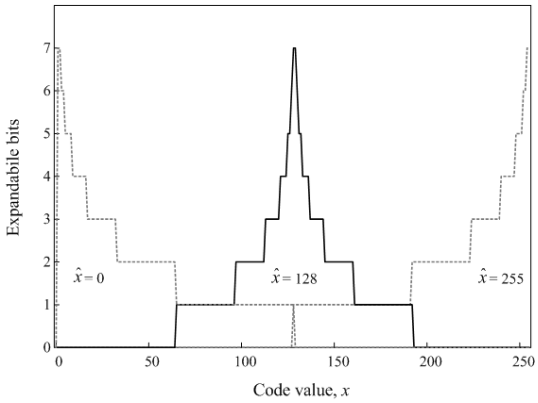


그림 4. 워터마크 비트가 전부 1일 때 $w=\{1\}_1^{2n-1}$, 예측계수 $\hat{x}$=0, 128, 255 상에서 x의 확장 가능 비트수

Fig. 4. In case that all of watermark bits are one; $w=\{1\}_1^{2n-1}$, expandable bits of x on estimated value $\hat{x}$=0, 128, 255.

계수 x 내에 은닉될 비트수를 보여준다. 최대 은닉 비트수 $k_{\max}$는 $2n-1=7$이다. 은닉 비트수가 증가할수록 확장 가능영역은 기하급수적으로 좁아지며, $\hat{x}$가 0 또는 255에 가까울수록 은닉될 비트수가 작아진다.

4. LS 기반 부호계수 예측

가역 영상 워터마킹에서는 예측오차를 줄이기 위하여 JPEG-LS 표준화에 사용된 MED(median edge detector)[16], CALIC(context-based, adaptive, lossless image coding)에 사용되는 GAP(gradient-adjusted predictor)와 SGAP(simplified GAP) [17] 등의 고정 예측기와 LS(least square)[18]의 적응 예측기 등 다양한 예측 방법들이 제시되었다. 인접 계수 간의 상관도가 낮은 부호계수 서열에서는 MED, GAP 등과 같은 고정 예측기보다 LS의 적응 예측기가 적합하다.

그림 5는 부호차수 n이 4일 때, 'AE017199', 'CP000473.1' 서열의 부호계수 및 부호계수 히스토그램을 보여준다. 부호계수 히스토그램은 부호차수에 따라 확장 및 축소가 되나, 서열에 따라 정형화된 분포를 가지지 않는다. 즉, 'AE017199' 서열은 네 영역을 제외한 나머지 영역에 골고루 분포되며, 'CP000473.1' 서열은 백색잡음과 같이 전체적인 영역에 골고루 분포된다. 또한 부호계수 서열은 랜덤 형태로 나타나며, 인접한 계수 간의 상관도가 매우 낮다. 따라서 부호계수 예측 오차를 줄이기 위하여 제안한 방법에서는 Dragoi 등의 지역 LS 예측[18]기반으로 부호계수를 예측한다.

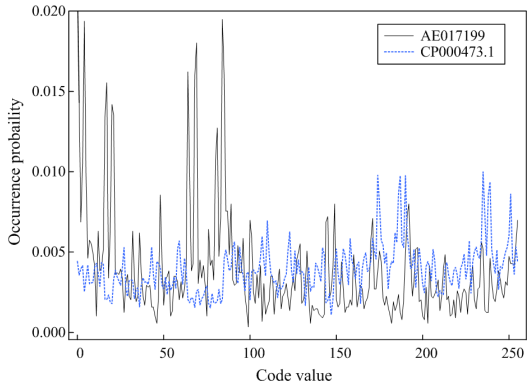


그림 5. 부호차수 n=4일 때, 'AE017199', 'CP000473.1' 서열의 인접 부호계수 차이 히스토그램

Fig. 5. Difference histogram of adjacent code values for 'AE017199', 'CP000473.1' sequences in code order (a) n=3 and (b) n=4.

현재 부호계수 x_i 예측을 위한 p개 부호계수 열벡터가 $\mathbf{x}_i=(x_{i-1},\cdots x_{i-p})$이고, p개 변수 열벡터 $\mathbf{b}=(\beta_1,\cdots,\beta_p)$라 한다. 이 때 p를 예측 차수라 한다. $\mathbf{x}_i$가 관측되었을 때, x_i의 예측계수 $\hat{x}_i$는 선형 회귀(linear regression) 함수 $f_\beta(\mathbf{x})$에 의하여

$$\hat{x}_i=f_\beta(\mathbf{x}_i)=\sum_{j=1}^{p}\beta_j x_{i-j}=\mathbf{x}_i\mathbf{b}' \tag{8}$$

와 같이 정의된다. 임의의 은닉 영역 내 전체 부호계수 열벡터 $\mathbf{y}=(x_1,\cdots,x_N)$이고, N 관측된 이전 부호계수들의 $N\times p$ 행렬 $\mathbf{X}=(\mathbf{x}_1',\cdots,\mathbf{x}_N')$라 할 때, LS 예측은 $\mathbf{y}'$와 $\mathbf{X}\mathbf{b}'$와의 제곱 거리 $\|\mathbf{y}'-\mathbf{X}\mathbf{b}'\|^2=(\mathbf{y}'-\mathbf{X}\mathbf{b}')'(\mathbf{y}'-\mathbf{X}\mathbf{b}')$가 최소가 되는 변수 $\mathbf{b}$를

$$\mathbf{b}=(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}' \tag{9}$$

와 같이 얻는다. 제안한 방법에서는 전체 은닉 영역에 대한 전역 예측보다 은닉 영역별 지역 예측을 통하여 부호계수를 예측한다. 따라서 복호 과정에서는 DNA 서열의 은닉 영역 개수 $|\Gamma(n)|$별 변수 $\mathbf{b}$인 $|\Gamma(n)|\times\mathbf{b}$의 부가정보가 필요하다.

부호계수 예측으로 인접계수 $\hat{x}_i=x_{i-1}$ 또는 평균 예측 $\hat{x}_i=\sum_{j=1}^{p}x_{i-j}/p$이 가능하다. 그림 6은

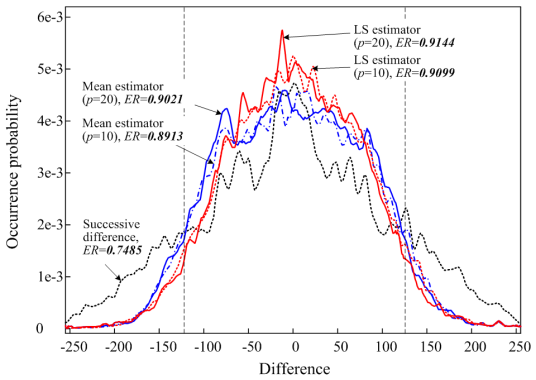


그림 6. 부호차수 n=4일 때, 'AE017199', 'CP000473.1' 서열의 LS 예측, 평균 예측, 및 인접계수와의 평균 오차 히스토그램 (p : 예측 차수(예측에 사용되는 인접 계수 개수), ER(expandible region)는 확장영역 발생확률)

Fig. 6. Histogram of prediction error of LS predictor, mean predictor, and adjacent code value for 'AE017199', 'CP000473.1' sequences in code order n=4.

'AE017199', 'CP000473.1' 서열에 대하여 부호차수가 $n=4$일 때 인접계수, 평균 예측, 및 LS 예측에 대한 예측오차 히스토그램을 보여준다. 그림에서 ER는 확장영역 발생확률을 나타낸다. 인접계수 오차는 부호차수에 상관없이 74.8% 정도의 확장영역을 가진다. 평균 예측과 LS 예측은 부호차수 n=3일 때 다소 높은 ER을 가지며, 예측차수 p가 높을수록 ER이 높아진다. 특히 n=3이고, p=20일 때, LS 예측이 가장 높은 91.6% 정도의 확장영역을 가진다. 즉, n=3일 때, LS의 예측차수 p가 높을수록 삽입 용량이 커짐을 알 수 있다.

영상의 예측 오차 히스토그램은 라플라시안 분포로 모델링되나, 부호계수의 LS 예측 오차 히스토그램은 결과에서와 같이 정규 분포로 모델링된다. n=3, p=10일 때, (μ,σ)=(0,20), n=3, p=20일 때 (μ,σ)=(0,19) 분포로 근사화된다. n=4, p=10일 때, (μ,σ)=(0,80), n=4, p=20일 때 (μ,σ)=(0,76) 분포로 근사화된다.

5. 가역 워터마크 삽입 과정

부호차수 n과 예측차수 p가 주어졌을 때, 은닉 영역별로 LS 예측 변수 $\mathbf{b}$를 구한 다음, $i>p$인 부호계수 x_i는 $\mathbf{b}$에 의한 LS 예측, $i \leqq p$인 부호계수는 평균 예측에 의하여 $\hat{x}_i$가 구하여진다.

$$\hat{x}_i = \begin{cases} \sum_{j=1}^{p} \beta_j x_{i-j}, & \text{if } i > p \\ \sum_{j=1}^{i-1} \frac{x_{i-j}}{i-1}, & \text{if } 1 < i \leqq p \\ 0, & \text{if } i = 1 \end{cases} \tag{10}$$

예측오차 $d_i = x_i - \hat{x}_i$의 확장 조건에 따라 은닉 비트수 k_i ($0 \leqq k_i \leqq 2n-1$가 결정된 후, 부호계수 x_i에 k_i 비트 $\{w_l\}_{l=1}^{k_i}$가

$$x_i' = \hat{x}_i + 2^{k_i} d_i + \alpha(k_i) \tag{11}$$

$$\text{where } \alpha(k_i) = sgn(d_i) \sum_{l=1}^{k_i} 2^{l-1} w_l \text{ and}$$

$$x_i' \notin \mathbf{Z}^c \text{ and } x_{i-1}'(n-1,n) \| x_i'(1,2) \notin \mathbf{Z}^c$$

와 같이 은닉한다. 은닉된 부호계수 x_i'가 허위개시코돈 테이블 $\mathbf{Z}^c$에 포함되거나, 이전 부호계수 x_{i-1}' 간 허위개시코돈을 포함할 경우 은닉 비트수 k_i를 하나 감소한 다음, k_i가 0일 때까지 위의 과정을 반복한다. 이와 같은 방법에 의하여 모든 은닉 영역의 부호계수에 다중비트를 은닉한 후, 워터마크된 영역 $\Gamma'(n)$을 얻는다. k_i가 0인 것은, 예측오차의 비은닉 영역에 해당되거나, 허위개시코돈 발생되는 경우를 나타낸다.

부호계수별 은닉 비트수 $\mathbf{K}=\{k_i\}$와 은닉 영역별 예측변수 $\mathbf{b}$는 워터마크 추출 및 원본 서열 복원에 필요한 부가정보이다. 부가정보는 허위개시코돈 발생되지 않고, 또 다른 부가정보 생성없이 워터마크된 영역 $\Gamma'(n)$에 포함되어 전송되어야 한다. 제안한 방법에서는 은닉 비트수 $\mathbf{K}$와 예측변수 $\mathbf{b}$와 $\Gamma'(n)$ 내 2bit 엠기 이진수의 LSB 비트 $\mathbf{B}$를 산술 부호화(arithmetic coding)에 의하여 무손실 압축하여, 압축 비트열 $\mathbf{C}=\{c_j\}$을 생성한다. 압축 비트 c_j는 4-문자 엠기의 이진수 b_i'의 LSB에

$$b_i'' = (b_i' \gg 1) \ll 1 + c_j \tag{12}$$

$$\text{if } b_{i-2}'' \neq \text{'A' and } b_{i-1}'' \neq \text{'T'}$$

와 같이 차례로 치환된다. 이 때, 이전 은닉된 두 개의 엠기 (b_{i-2}', b_{i-1}')가 "AT"인 경우, 현재 엠기가 b_i'='G'이면 b_i'를 'A','T','C' 중에 하나로 치환하고, $b_i' \neq$'G'이면 은닉 과정을 생략한다. 최종적으로 압축열 $\mathbf{C}$을 포함하는 은닉 영역 $\Gamma''(n)$ 내에 "AT" 엠기열은 다음 엠기에 압축 비트가 포함되지 않음을 직접적으로 나타내는 마커로 수행된다. 최종 부가정보 및 워터마크가 은닉된 비부호 영역 $\Gamma''(n)$을 가지는 DNA 서열 $\mathbf{D}' = \mathbf{D}'^{nc} + \mathbf{D}^c$, , $\mathbf{D}'^{nc} = \Gamma''(n) + \Gamma^c(n)$이 전송된다.

6. 워터마크 검출 및 DNA 서열 복원 과정

복호 과정에서는 그림 1(b)에서와 같이 먼저 전송된 DNA 서열 $\mathbf{D}'$의 비부호 영역 $\Gamma''(n)$ 상에서 "AT" 다음에 오는 엠기를 제외한 모든 엠기들의 LSB로부터 부가정보 압축열 $\mathbf{C}$의 은닉 비트수 $\mathbf{K}$, 예측변수 $\mathbf{b}$와 엠기 LSB 비트 $\mathbf{B}$를 얻는다. $\Gamma''(n)$의 엠기 LSB 비트에 $\mathbf{B}$가 치환된 $\Gamma'(n)$을 부호차수 n에 의하여 부호서열 X'를 얻는다. X' 내 모든 부호계수로부터 은닉 비트수 $\mathbf{K}$와 예측변수 $\mathbf{b}$에 의하여 워터마크를 추출하고, 원본 부호계수를 복원한다. 예를 들어, 은닉 비트수 $k_i>0$인 임의의 부호계수 x_i'가 주어졌을 때, 이전 복원된 부호계수$(x_{i-1},\cdots,x_{i-p})$로부터 예측계수 $\hat{x}_i$가 구한 다음,

예측오차 $d_i = x_i^{'} - \hat{x}_i$로부터 워터마크 k_i 비트가 $w_l = ((x_i^{'} - \hat{x}_i) \gg (l-1))\%2, \quad \text{for } l = 1, \cdots, k_i$에 의하여 추출된다. 그리고 원본 부호계수 x_i는 예측오차 d_i를 k_i 비트 쉬프팅에 의하여 $x_i = \hat{x}_i + ((x_i^{'} - \hat{x}_i) \gg k_i)$와 같이 복원된다.

Ⅳ. 실험 결과

1. 실험 환경

영상 가역 워터마킹의 성능 평가에서는 용량(bpp; bit per pixel)에 대한 PSNR이 사용되나, DNA 가역 워터마킹에서는 PSNR의 화질 척도 대신 생물학적 기능 변경이 사용된다. 제안한 방법에서는 비부호 영역에 허위개시코돈 발생되지 않도록 가역 워터마크를 은닉하므로 생물학적 기능 변경이 없다. 따라서 본 실험에서는 제안한 LS 예측오차 확장 방법 (LS-PE)과 평균 예측오차 확장 방법 (Mean-PE), 기존의 연속 부호계수 차이 방법[15], Chen 방법[12], 및 Huang 방법[13]의 워터마크 용량 bpn_W, 압축 부가정보량 bpn_{Extra}, 염기 변화율 $e(n)$, 및 허위개시코돈 발생 확률에 대하여 비교 분석하였다. 여기서 염기 변화율 $e(n)$은 워터마크에 의하여 변경된 염기의 비율을 나타내는 것으로, 원본 은닉 대상 서열 $\Gamma^{\mathrm{nc}}(n)$과 워터마크된 서열 $\Gamma'^{\mathrm{nc}}(n)$ 일 때,

$$e(n) = \frac{1}{|\Gamma^{\mathrm{nc}}(n)|\,|\mathrm{D}_k^{\mathrm{nc}}|} \sum_{k=1}^{|\Gamma^{\mathrm{nc}}(n)|} \sum_{i=1}^{|\mathrm{D}_k^{\mathrm{nc}}|} e_{ki} \tag{13}$$

$$\text{where } e_{ki} = \begin{cases} 1, \text{if } b_{ik} \neq b_{ik}^{'} \\ 0, \text{if } b_{ik} = b_{ik}^{'} \end{cases}$$

와 같이 정의된다. 임의의 워터마크에 의하여 염기가 변경될 확률이 전체적으로 균등분포를 가진다고 가정할 때, 염기 변화율은 3/4=0.75에 가깝다.

제안한 방법의 워터마크 용량과 부가정보량은 부호차수 n과 예측차수 p에 의하여 결정된다. 따라서 본 실험에서는 부가정보량과 최대 워터마크 용량을 가지는 LS 기반 PE 방법의 예측차수 p를 실험적으로 선택한 후, 각 방법들의 용량을 부호차수별로 비교 분석하였다. 이 때 동등한 성능 평가를 위하여 Chen 방법($|w|$=2)과 Huang 방법(t=2)에서는 워터마크 용량이 제일 높을 때의 변수를 사용하였다.

본 실험에서는 NCBI GenBank에서 제공된 15개의 DNA 서열을 이용하였으며, 이들 서열들의 비부호 영역들은 다양한 염기 개수를 가지며, 매우 작은 염기 수로 구성된 영역들은 제안한 영역 선택 과정에 의하여 은닉 대상에서 제외된다.

2. 워터마크 용량과 부가정보량 분석

제안한 방법의 워터마크 용량은 부호차수 n, 예측차수 p에 의하여 영향을 받는다. n과 p가 주어졌을 때, 은닉 영역 $\Gamma(n) = \{\mathrm{D}_i\}_{i=1}^{|\Gamma(n)|}$내에 은닉되는 워터마크 비트수는 각 영역 내 부호계수 은닉 비트수 K의 합에 해당된다. 따라서 염기 당 비트수 bpn(bit per base) $bpn_W(n,p)$은

$$bpn_W(n,p) = \frac{1}{|\Gamma(n)|} \sum_{i=1}^{|\Gamma(n)|} \left(\frac{1}{N_i} \sum_{j=1}^{N_i} k_j\right) [\text{bit/base}] \tag{14}$$

$$\text{where } N_i = \lfloor |\mathrm{D}_i|/n \rfloor \text{ and } 0 \leqq k_j \leqq 2n-1$$

와 같다. $|\Gamma(n)|$는 은닉 영역의 개수이고, N_i는 영역 D_i 내 부호계수의 개수를 나타낸다.

부가정보 압축열 C을 은닉하기 위한 LSB 치환 가능 비트량을 Φ라 할 때, Φ는 치환 과정에서 허위개시코돈에 의하여 생략되는 염기의 개수에 의하여 결정된다. 최대 Φ는 $\Gamma'(n)$ 내의 총 염기 개수 $\sum_{i=1}^{|\Gamma(n)|} |\mathrm{D}_i|$와 동일하다. 부가정보 압축열 C 길이는 치환 가능 비트량 Φ 보다 작아야 하므로, 은닉 비트수 K, 예측변수 b 및 2bit 염기의 LSB B의 부가 정보량이 작거나, 압축효율이 높은 알고리즘이 필요하다. 임의의 워터마크된 영역 $\mathrm{D}_i^{'}$ ($\in \Gamma'(n)$)이 주어졌을 때, B는 $|\mathrm{D}_i|$ 비트로 구성되며, 은닉 비트수 K는 $N_i \lceil \log_2 2n \rceil$ 비트로 표현되고, 은닉영역별 예측변수 b는 예측차수 p개의 32비트 부동소수점로 표현된다. 따라서 $\Gamma'(n)$을 위한 부가정보는

$$Extra_{\mathrm{PE}}(n,p) = \sum_{i=1}^{|\Gamma(n)|} (N_i \lceil \log_2 2n \rceil + |\mathrm{D}_i| + 32p)$$

[bit]이다. 부가정보 압축열 C가 $\rho \times Extra_{\mathrm{PE}}(n,p)$ 라 할 때, $\rho \times Extra_{\mathrm{PE}}(n,p) < \Phi \leqq \sum_{i=1}^{|\Gamma(n)|} |D_i|$이 되도록 압축이 수행된다. 따라서 염기 당 부가정보 비트수 $bpn_{Extra}(n,p)$는

$$bpn_{Extra}(n,p) = \frac{1}{|\Gamma(n)|}\sum_{i=1}^{|\Gamma(n)|}\rho(N_i\lceil \log_2 2n \rceil + |\mathbf{D}_i| + 32p) \text{[bit/base]} \quad (15)$$

와 같다.

제안한 방법은 부호차수 n가 주어졌을 때, 예측차수 p가 커질수록 확장 발생 확률이 높아짐을 알 수 있었다. 부호차수 n이 [2,3,4]로 주어졌을 때, LS 예측과 평균 예측 기반 PE 방법의 워터마크 bpn bpn_W는 그림 7에서와 같다. 즉, 부호차수 n이 작고, p가 커질수록 bpn_W이 커짐을 볼 수 있다. 또한 평균 예측은 예측차수 p가 3일 때부터 워터마크 bpn이 유지됨을 알 수 있다. n, p에 따라 평균 예측보다 LS 예측이 평균적으로 0.02 bpn이 보다 높다. 부호차수 n이 2일 때, 예측차수 p가 20일 때부터 LS 기반 PE 방법의 워터마크 bpn은 0.413~0.419 정도 미세하게 증가됨을 볼 수 있다.

예측차수 p가 클수록 워터마크 bpn 뿐만 아니라 부가정보량과 염기변화율이 커지므로, 이를 고려하여 예측차수 p가 결정되어야 한다. 그림 8은 부호차수 n=2,3,4일 때 LS 기반 PE 방법의 압축된 부가정보량 대비 워터마크 bpn과 염기 변화율 대비 워터마크 bpn를 보여주고 있다. 그림 8(a)에서와 같이 부호차수가 낮을수록 높은 워터마크 bpn과 낮은 부가정보량을 가지나, 예측차수가 클수록 부가정보량이 커짐을 볼 수 있다. 부가정보량이 1bpn에 가까울수록 은닉 대상 영역의 치

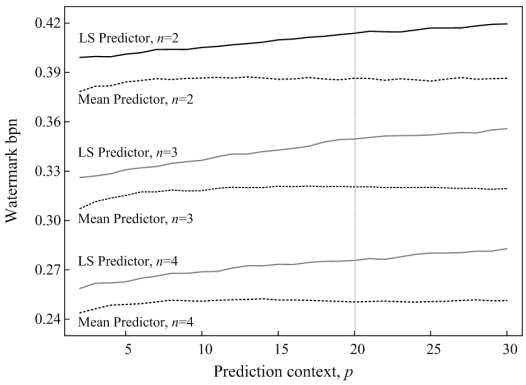
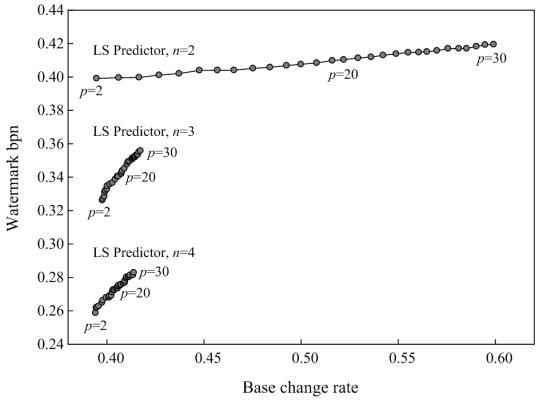
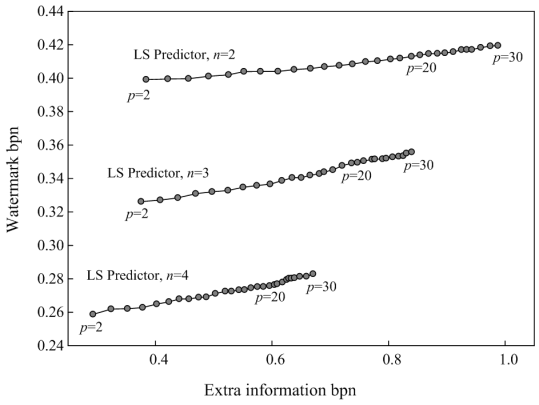


그림 7. 모든 서열에 대하여 부호차수 $n(\in[2,4])$이 주어졌을 때 예측차수 $p(\in[2,30])$에 대한 평균 워터마크 bpn

Fig. 7. Average watermark bpn on code order $n\in[2,4]$ and prediction order $p\in[2,30]$ for all test sequences.



(a)



(b)

그림 8. 부호차수 n=2,3,4일 때 LS 기반 PE 방법의 (a) 부가정보량 vs 워터마크 bpn와 (b) 염기 변화율 vs 워터마크 bpn

Fig. 8. (a) Extra information bpn vs watermark bpn and (b) base change rate vs watermark bpn of our LS-PE method in code order n=2,3,4.

환 가능 비트량 Φ에 가깝거나 보다 커질 수 있다. 그림 8(b)에서와 같이 부호차수가 낮을수록 낮은 염기변화율에 높은 워터마크 bpn은 가짐을 볼 수 있다. 염기변화율이 0.5일 때, 은닉 대상 염기들 중 50%가 자신 이외의 다른 세 염기들 중 하나로 변경됨을 나타낸다. $n=2$에서 $p=20$일 때, 워터마크 bpn=0.414, 부가정보량 bpn=0.854, 염기변화율=0.548이고, $p=30$일 때, 워터마크 bpn=0.419, 부가정보량 bpn=0.988, 염기변화율=0.599이다. 본 실험에서는 부가정보량 및 염기변화율을 고려하면서 높은 워터마크 bpn를 가지는 예측차수 p를 20으로 설정하였다.

3. 성능 비교 : 워터마크 bpn, 부가정보량 bpn, 염기변화율

본 실험에서는 제안한 LS-PE 방법과 Mean-PE 방법의 예측차수 p를 20,30으로 설정하고, 부호차수 n이 2에서 6까지 가변하여 기존 방법과의 워터마크 bpn bpn_W, 부가정보량 bpn bpn_{Extra}, 및 염기변화율과 용량 효율(bpn_W/bpn_{Extra})을 비교하였다. 기존 방법들은 부호차수 n에 상관없으므로, 하나의 결과를 가진다. 이에 대한 결과를 그림 9에 나타내었다.

워터마크 bpn bpn_W 결과인 그림 9(a)를 살펴보면, LS-PE 방법과 Mean-PE 방법은 부호차수 n이 증가할수록 bpn_W이 감소된다. LS-PE 방법의 bpn_W이 (n,p)=(2,30)일 때 0.419bpn으로 가장 높으며, (n,p)=(2,20)일 때 0.413bpn이다. 다음 Mean-PE 방법은 (n,p)=(2,30),(2,20)일 때, 각각 0.386bpn, 0.377bpn이다. (n,p)이 가변될 때, LS-PE 방법이 Mean-PE 방법보다 0.022~0.049bpn 정도 높음을 확인할 수 있다. 기존 연속 부호계수 차이 방법, Chen 방법과 Huang 방법은 부호차수에 상관없이 각각 0.235bpn, 0.109bpn과 0.026bpn으로 낮게 나타났다.

워터마크 추출에 필요한 부가데이터 bpn bpn_{Extra}결과인 그림 9(b)를 살펴보면, LS-PE 방법은 (n,p)=

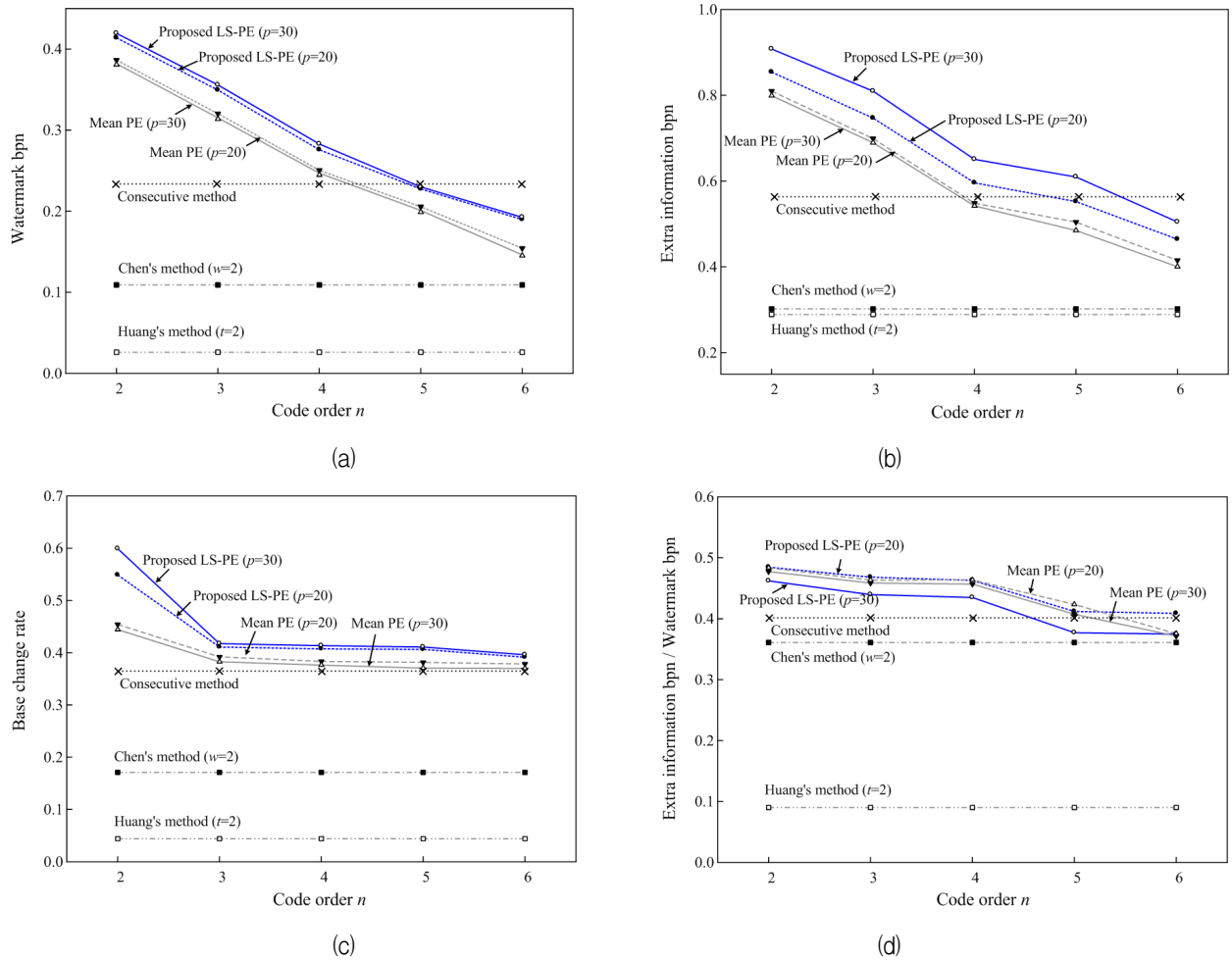


그림 9. 제안한 LS 기반 PE 방법과 평균 예측오차 확장 방법 (p=20,30)과 기존 Chen 방법 및 Huang 방법에 대한 (a) 워터마크 bpn, (b) 부가정보량 bpn, (c) 용량효율 (부가정보량 bpn/워터마크 bpn), (d) 염기변화율 (기존 방법은 부호차수 n에 상관없는 결과임)

Fig. 9. Results of our LS-PE method and mean-PE method with (p=20,30), Chen's method, and Huang's method; (a) watermark bpn, (b) extra information bpn, (c) capacity efficiency (extra information/watermark), and (d) base change rate (Chen's method and Huang's method are not depend on code order n).

(2,30),(2,20)일 때, 각각 0.908bpn과 0.854bpn으로 높게 나타났다. (n,p)가 가변될 때, LS-PE 방법이 Mean-PE 방법보다 0.044~0.125bpn 정도 높은 bpn_{Extra}를 가진다. 이는 워터마크 bpn이 많을수록 추출위한 부가데이터가 많이 필요함을 나타낸다. 기존의 연속 부호계수 차이 방법 및 Chen과 Huang 방법들은 워터마크 bpn이 낮으므로, 0.582bpn, 0.302bpn과 0.289bpn의 낮은 부가데이터 bpn_{Extra}가 필요하다.

부가데이터 bpn 대비 워터마크 bpn 비율을 나타내는 용량 효율 (bpn_W/bpn_{Extra})의 결과인 그림 9(c)를 살펴보면, LS-PE 방법과 Mean-PE 방법이 0.371~0.485 정도로 부가데이터 1비트당 0.371~0.485비트의 워터마크를 은닉할 수 있다. 그러나 기존 연속 부호계수 차이 방법 및 Chen과 Huang 방법은 0.408, 0.361과 0.090의 낮은 용량 효율을 가진다.

워터마크에 의한 염기변화율의 결과인 그림 9(d)를 살펴보면, 워터마크 bpn이 높을수록 염기변화율이 높게 나타난다. 즉, LS-PE 방법은 (n,p)=(2,30),(2,20)일 때, 각각 0.599와 0.549로 높게 나타났으며, 부호차수 $n>2$일 때 약 0.40에 근접하게 나타났다. Mean-PE 방법도 $n=2$일 때, 약 0.45 정도이고, $n>2$일 때 약 0.38에 근접하게 나타났다. 이와 반대로 기존 연속 부호계수 차이 방법 및 Chen 방법과 Huang 방법은 낮은 워터마크 bpn으로 0.370, 0.171과 0.044로 낮게 나타났다. 비부호 DNA 서열은 Junk 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. 따라서 아미노산 변화율은 0이다. 만약 비부호 DNA 서열에서 발현 인자 조절과 같은 주요한 염기 서열이 알려졌을 경우, 허위개시 코돈 방지와 같은 방법으로 이들 서열이 변경되지 않도록 한다.

4. 허위개시코돈 확률

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

$$p_f = P(\mathrm{b}_i\mathrm{b}_{i+1}\mathrm{b}_{i+2} = "\mathrm{ATG}"|\mathbf{D}'^{nc}) \tag{16}$$

을 허위 개시코돈 발생 확률이라 한다. 본 실험에서는 모든 테스트 DNA 서열의 1,000번 반복 수행하여 발생된 허위개시코돈 확률 p_f를 구하였다.

제안한 방법은 워터마크 은닉 과정과 부가 데이터의 LSB 치환 과정에서 허위 개시코돈 방지를 위하여 비교 탐색 과정을 수행한다. 따라서 모든 실험 상에서 제안한 LS-PE 방법과 Mean-PE 방법이 허위개시코돈이 발생되지 않음을 확인하였다. 그러나 기존의 Chen 방법과 Huang 방법은 허위개시코돈 발생에 대하여 고려하지 않으며, 실험의 워터마크 은닉 과정에서 허위 개시코돈이 발생되었다. 즉, Chen의 방법은 10^{-4} 염기 당 하나의 허위개시코돈이 발생되며, Huang의 방법은 5.78×10^{-5} 염기 당 하나의 허위개시코돈이 발생됨을 확인하였다.

V. 결 론

본 논문에서는 비부호 영역을 이용한 최소자승 예측오차 확장 기반의 가역 DNA 워터마킹 기법을 제안하였다. 제안한 방법은 부호차수에 의한 염기서열 부호화 및 예측차수에 의한 최소자승 오차 확장 (LS-PE)에 의하여 부호계수별 다중비트 은닉이 가능하고, 부호계수의 인트라 부호와 인터 부호 검색에 의하여 비부호 영역이 부호 영역으로 인식될 수 있는 허위개시코돈 생성을 방지한다. 실험 결과로부터 제안한 LS-PE 방법은 부호차수 및 예측차수가 2, 30일 때, Mean-PE 방법, 연속 부호계수 차이 방법, Chen 방법, Huang 방법보다 약 0.033~0.395bpn 정도 높은 워터마크 bpn를 가지며, 부가정보량 대비 워터마크량이 약 0.46 정도로 기존 방법보다 0.101~0.372 정도 높음을 확인하였다. 또한 LS-PE 방법과 Mean-PE 방법은 허위개시코돈이 발생되지 않았으나, 기존 방법들은 10^{-4}~5.78×10^{-5} 확률로 허위개시코돈이 발생됨을 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, E. Birney, "Towards practical high-capacity, low-maintenance information storage in synthesized DNA," *Nature*, vol. 494, pp. 77-80, Feb. 2013.

[3] C. T. Clelland, V. Risca, C. Bancroft, “Hiding messages in DNA microdots,” *Nature*, vol. 399, pp. 533-534, June 1999.
[4] M. Borda and O. Tornea, “DNA secret writing techniques,” *8th International Conference on Communications (COMM)*, pp. 451-456, June 2010.
[5] S.-H. Lee and K.-R. Kwon, “DNA Information Hiding Method for DNA Data Storage,“ *Journal of The Institute of Electronics and Information Engineers*, vol. 51, no. 10, pp. 118-127, Oct. 2014.
[6] D. Heider and A. Barnekow, “DNA-based watermarks using the DNA-Crypt algorithm,” *BMC Bioinformatics*, vol. 8, no. 176, May 2007.
[7] D. Heider and A. Barnekow, “DNA Watermarks - A proof of concept,” *BMC Bioinformatics*, vol. 9, no. 40, April 2008.
[8] 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.
[9] S.-H. Lee, “DNA sequence watermarking based on random circular angle,“ *Digital Signal Processing*, vol. 25, pp. 173-189, Feb. 2014.
[10] S.-H. Lee, “DWT based coding DNA watermarking for DNA copyright protection,“ *Information Sciences*, vol. 273, pp. 263-286, July, 2014.
[11] S.-H. Lee, S.-G. Kwon, 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, no. 10, pp. 91-101, Oct. 2012.
[12] 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.
[13] 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, June 2014.
[14] 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.
[15] S.-H. Lee and K.-R. Kwon, “Consecutive Difference Expansion Based Reversible DNA Watermarking,“ *Journal of The Institute of Electronics and Information Engineers*, To be published on July 2015.
[16] D.M. Thodi et al. “Expansion Embedding Techniques for Reversible Watermarking,” *IEEE Trans. on Image Processing*, vol. 16, no. 3, March 2007.
[17] D. Coltuc, “Improved Embedding for Prediction-Based Reversible Watermarking,“ *IEEE Transactions on Information Forensics and Security*, vol. 6, issue 3, pp. 873-882, Sept. 2011.
[18] I.-C. Dragoi and D. Coltuc, “Local-Prediction-Based Difference Expansion Reversible Watermarking,“ *IEEE Transactions on Image Processing*, vol. 23, issue 4, pp. 1779-1790, April 2014.
[19] W.-J. Kim, P.-H. Kim, J.-H. Le, K.-H. Jung, and K.-Y. Yoo, “Reversible Data Hiding Method Based on Min/Max in 2x2 Sub-blocks,“ *Journal of The Institute of Electronics and Information Engineers*, vol. 51, no. 4, pp. 69-75, April 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 : 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>