논문 2012-49-10-12

Lifting 기반 1D DWT 영역 상의 강인한 DNA 워터마킹

(A Robust DNA Watermarking in Lifting Based 1D DWT Domain)

이 석 환*, 권 기 룡**, 권 성 근***

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

요 약

개인 유전정보 또는 대용량 DNA 저장 정보의 보호와 GMO(Genetically Modified Organism) 저작권 보호를 위하여 DNA 서열 워터마킹 연구가 필요하다. 기존 멀티미디어 데이터 워터마킹에서는 강인성 및 비가시성에 대한 성능이 우수한 DCT, DWT, FMT(Fourer-Mellin transform) 등 주파수 기반으로 설계되어졌다. 그러나 부호 영역 서열의 주파수 기반 워터마킹은 아미노산 보존성을 유지하면서 변환 및 역변환을 수행하여야 하므로, 워터마크 삽입에 대한 상당한 제약을 가진다. 따라서 본 논문에서는 변이 강인성, 아미노산 보존성 및 보안성을 가지는 부호 영역 서열의 Lifting 기반 DWT 변환 계수를 이용한 워터마킹을 제안하며, 주파수 기반 DNA 서열 워터마킹에 대한 가능성을 제기한다. 실험 결과로부터 제안한 방법이 10%의 포인트 변이와 5%의 삽입 및 삭제 변이에 대한 강인성을 가지며, 아미노산 보존성 및 보안성을 가짐을 확인하였다.

Abstract

DNA watermarking have been interested for both the security of private genetic information or huge DNA storage information and the copyright protection of GMO. Multimedia watermarking has been mainly designed on the basis of frequency domain, such as DCT, DWT, FMT, and so on, for the robustness and invisibility. But a frequency domain watermarking for coding DNA sequence has a considerable constraint for embedding the watermark because transform and inverse transform must be performed without completely changing the amino acid sequence. This paper presents a coding sequence watermarking on lifting based DWT domain and brings up the availability of frequency domain watermarking for DNA sequence. From experimental results, we verified that the proposed scheme has the robustness to until a combination of 10% point mutations, 5% insertion and deletion mutations and also the amino preservation and the security.

Keywords : 바이오인포매틱스(Bioinformatics), DNA 워터마킹(DNA watermarking), 부호 DNA 서열(Coding DNA sequence), Lifting 기반 DWT(Lifting based DNA), 유전 정보 보호(Genetic information security), 변이 강인성(Mutation robustness)

Ⅰ. 서 론

BIT(Biology Information Technology) 발달에 따라 개인 유전 정보의 보호와 대용량 정보 저장을 위한 DNA 저장 인식으로 인한 DNA 정보의 보호 필요성이 제기되고 있다. 이와 같은 필요성에 따라 DNA 암호화[1~3], DNA 스테가노그라피[4~8], DNA 워터마킹[9~15]에 대한 연구가 진행되어지고 있다. DNA 암호화는 생물학적인 DNA 암호 및 복호 방법을 제시한 것으로, PCR(Polymerase chain reaction)과 DNA 칩 기반의 암

* 정회원, 동명대학교 정보보호학과
(Dept. of Information Security, Tongmyong University)
** 정회원, 부경대학교 IT융합응용공학과
(Dept. of IT Convergence and Application Engineering, Pukyong National University)
*** 정회원-교신저자, 경일대학교 전자공학과
(Dept. of Electronics Engineering, KyungIl University)
※ 이 논문은 2011년도 정부(교육과학기술부)의 재원으로 한국연구재단의 기초연구사업 지원을 받아 수행된 것입니다. (KRF-2011-0023118)
접수일자:2012년6월27일, 수정완료일:2012년10월8일

호화가 대표적이다. PCR 기반 암호화[1~2]는 PCR 프라이머 쌍(Primer pair)을 암호키로 사용한 것으로, 프라이머 쌍을 모를 경우, 대상 DNA를 복호할 수 없도록 하는 것이다. 그리고 DNA 칩 기반 암호화[3]는 특정 프로브들을 암호키로 사용하여 DNA 칩 안에 암호 데이터를 삽입하는 것으로 실험 환경이 다르거나 복호키가 없을 경우 DNA 칩으로부터 데이터를 복호할 수 없도록 하는 것이다. 그러나 DNA 암호화는 실제 구현의 어려움으로 인하여 기존 암호 기술을 대체할 수 없으나, 새로운 생물학적 기반 암호 기술을 제시할 수 있는 분야로 인식되고 있다. DNA 스테가노그라피는 DNA에 정보를 은닉하는 기술[4~8]로, DNA에 대용량 정보 저장, 또는 DNA 서명 및 인식에 매우 유용하다.

DNA 암호화 및 스테가노그라피는 실험 환경 또는 DNA 변이 등과 같은 경우에 DNA 또는 정보를 복원할 수 없으므로, 강인성이 요구되는 응용에서는 적합하지 못하다. 이와 같은 DNA 암호화 및 스테가노그라피의 문제점을 해결하기 위하여 DNA 워터마킹 기법들이 일부 연구자들에 의하여 제안되어지고 있다. Heider 등은 4개의 염기서열에 따라 이진 정보를 삽입하는 DNA-Crypt 알고리즘과 8/4 해밍 코드 또는 WDH 코드의 변이 정정 부호기를 결합한 방법을 제안하였다[12]. 이들은 변이 정정 부호기 기반의 DNA-Crypt 알고리즘을 비부호 영역[11]뿐만 아니라 코돈 중복성 기반의 부호 영역[13]에 삽입하였으며, 또한 미토콘드리아에 적합한 DNA-Crypt 알고리즘을 제안하였다[14]. 이 방법은 스테가노그라픽의 LSB (Least Significant Bit) 치환 방법과 같은 단순 염기 서열 치환에 오류 정정 부호 기법을 결합한 것이다. 멀티미디어 워터마킹[16~19]에서는 강인성과 보안성 문제점으로 인하여 LSB 치환 방법을 사용하지 않으며, DCT, DWT, SIFT 등 다양한 변환 영역 또는 기하학적 신호 처리 상에 워터마크를 삽입한다. 특히 DWT는 DNA 주파수 분석, 부호 영역 식별 등의 DNA 신호 해석에 많이 응용되고 있다. 그러나 워터마크된 주파수 계수의 역변환에 의하여 유전 정보가 쉽게 변경될 수 있다. 즉, 아미노산 보존 조건 하에서 주파수 변환 기반의 워터마크 삽입이 가능하여야 한다. 이와 같은 제한 조건으로 인하여 주파수 변환 기반의 DNA 워터마킹에 대한 연구가 이루어지지 않고 있다.

본 논문에서는 부호 DNA 서열의 DWT 변환 영역 상에 워터마크를 삽입하는 방법에 대하여 논의하고, 변이 강인성, 아미노산 보존성, 및 워터마크 보안성에 대한 성능을 신호처리 관점에서 검증한다. DWT 기반의 DNA 신호 해석과는 달리 부호 DNA 워터마킹에서는 아미노산 보존성에 따라 DWT 변환 및 역변환 과정이 수행되어야 하므로, 워터마크 삽입이 용이하지 못하다. 즉, 워터마크된 DWT 계수의 역변환된 코돈 서열에서는 아미노산이 보존되어야 한다. 제안한 방법에서는 부호 영역의 전체 코돈 서열을 코돈 부-서열들로 분할한 후, 각 코돈 부-서열 단위로 코돈 중복성을 이용한 DWT 변환 및 역변환 과정을 수행한다. 이 때 각 코돈 부-서열은 워터마크 성분에 가장 적합한 DWT 계수 서열을 가지는 코돈 부-서열로 치환된다. 그리고 워터마크 보안성 향상을 위하여 코돈 부-서열 단위로 삽입 대상 계수를 랜덤하게 선택된다. 실험 결과로부터 제안한 방법이 DNA-Crypt 알고리즘에 비하여 10%의 치환 변이와 5%의 삽입 및 삭제 변이에서 높은 강인성과 보안성을 가지며, 아미노산 보존성을 유지함을 확인하였다.

본 논문의 구성은 다음과 같다. Ⅱ장에서는 DNA 워터마킹 프레임워크와 관련 연구를 살펴보며, Ⅲ장에서는 기반 부호 DNA 워터마킹 방법에 대하여 자세히 살펴본다. Ⅳ장에서는 제안한 방법에 대한 강인성, 보안성 및 유일성 평가 실험에 대하여 분석하며 마지막으로 Ⅴ장에서는 부호 DNA 워터마킹에서의 DWT 효율성에 대한 결론을 맺는다.

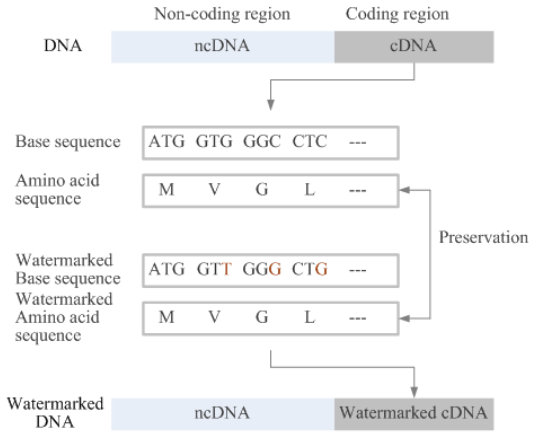
Ⅱ. 부호 DNA 서열 워터마킹

본 장에서는 부호 DNA 서열 상에 워터마크 삽입 과정에 대하여 간략히 살펴보기로 한다.

코돈은 그림 1(a)에서와 같은 유전 부호에 의하여 하나의 아미노산으로 대응된다[20~21]. 이 때 Methionine (M)과 Tryptophan (T)를 제외한 나머지 모든 아미노산들에 대해 복수의 코돈이 대응되는 코돈 중첩(Codon degeneracy)이 발생되며, 동일 아미노산 내에 다른 코돈들을 동의어 코돈(Synonymous codon)이라 한다. 부호 DNA 워터마킹에서는 아미노산 보존성을 위하여 삽입 대상 코돈들이 워터마크에 의하여 동의어 코돈들 중 하나로 변경되어야 한다. 그림 1(b)는 코돈 중첩에 의하여 워터마크된 부호 DNA 서열의 생성 예시를 보여준다. 이 과정을 살펴보면, 개시 및 종료 코돈을 제외한 나머지 코돈들 중 4중 중첩을 가지는 코돈들의 세 번째

AAA, AAG : K (Lys) AAT, AAC : N (Asn)	GAA, GAG : E (Glu) GAT, GAC : D (Asp)	TAA, TAG : Stop TAT, TAC : Y (Tyr)	CAA, CAG : Q (Gln) CAT, CAC : H (His)
AGA, AGG : R (Arg) AGT, AGC : S (Ser)	GGA, GGG, GGT, GGC : G (Gly)	TGA : Stop TGG : W (Trp) TGT, TGC : C (Cys)	CGA, CGG, CGT, CGC : R (Arg)
ATA : I (Ile) ATG : M Start ATT, ATC : I (Ile)	GTA, GTG, GTT, GTC : V (Val)	TTA, TTG : L (Leu) TTT, TTC : F (Phe)	CTA, CTG, CTT, CTC : L (Leu)
ACA, ACG, ACT, ACC : T (Thr)	GCA, GCG, GCT, GCC : A (Ala)	TCA, TCG, TCT, TCC : S (Ser)	CCA, CCG, CCT, CCC : P (Pro)

(a)



(b)

그림 1. (a) DNA 유전 부호
(b) 워터마크된 cDNA 생성 예시
Fig. 1. (a) DNA genetic coding and
(b) example of watermarked CDS generation.

염기에 2비트씩 삽입된다. 여기서 n중 중첩 코돈에는 n 비트가 삽입될 수 있다. 이와 같은 방법은 기존 DNA 스테가노그라피[4~8] 또는 DNA 워터마킹[9~15]에서 많이 사용되며 대표적인 방법으로 Heider 등의 DNA-Crypt 알고리즘[13]이 있다. DNA-Crypt에서는 4개의 염기 {A, C, G, T}에 대한 2비트를 할당한 다음, 4중 중첩 코돈의 마지막 염기에 2비트씩 삽입한다. 여기서 이 알고리즘에서는 변이 정정을 위한 8/4 Hamming 부호와 n-time WDH 부호를 변이율, 서열 길이, 시간 안정성 변수를 입력으로 한 퍼지 제어기를 통하여 적합한 변이 정정 부호를 선택한다. 그리고 AES, Blowfish, RSA 또는 OTP(One-Time pad)의 암호화 알고리즘에 의한 워터마크 이진 부호의 암호화가 가능하다. 그러나 암호화된 워터마크를 사용하더라도 삽입 위치가 공개되어 있으므로, 제3자에 의한 워터마크 제거가 용이하다.

DNA 신호 해석에서는 신호처리 용이를 위하여

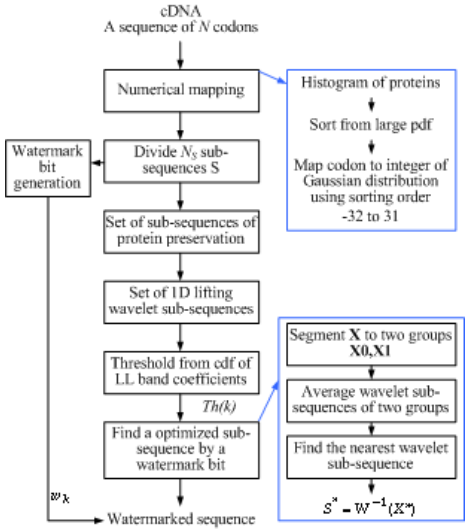


그림 2. 제안한 DNA 워터마크 삽입 과정
Fig. 2. Process of proposed DNA watermark embedding.

DNA 서열을 정수, 실수 또는 복소수 형태의 수치로 사상한다. 즉, 부호 DNA 워터마킹에서는 코돈 또는 염기를 임의의 수치로 사상한 다음, 수치 서열의 변환 계수 또는 신호처리에 의하여 워터마크를 삽입할 수 있다. Cristea[22~23]는 64개의 코돈과 종결 부위를 포함한 아미노산과의 최소 비단조 대응 조건 하에 최적의 염기 부호로 T=0, C=1, A=2, G=3으로 선택하여, 이에 따른 코돈 부호를 제시하였다. 본 논문에서는 Cristea의 정수 사상 기반으로 DWT 변환 및 역변환과 보안성을 고려한 코돈 정수 사상을 제안하며, 이에 따른 DWT 상에서의 워터마크 삽입 가능성을 제기한다.

Ⅲ. 제안한 DWT 기반 부호 DNA 서열 워터마킹

본 논문에서는 DWT 주파수 영역 상에서 적용이 가능한 부호 DNA 서열 워터마킹 기법을 제안한다. 제안한 워터마크 삽입 방법에서는 부호 DNA 서열 상에 코돈들의 수치 사상, 부-코돈 수치 서열의 DWT 변환, 및 워터마크 삽입 과정으로 구성되며, 추출 방법은 삽입 방법과 유사하다.

가. 워터마크 삽입

제안한 워터마크 삽입 과정은 그림 2에 자세히 나타내고 있다. 워터마크 비트는 N 길이의 부-코돈 수치 서열의 DWT 변환 계수에 삽입된다. DWT 변환을 위하

여 코돈들은 아미노산 히스토그램 분포에 따라 설정된 순위표에 의하여 수치 사상된다. 그리고 N 길이의 부-코돈 수치 서열은 동일 아미노산 서열을 가지는 DWT 변환 계수 집합 중에서 워터마크 비트 삽입에 가장 적합한 변환 계수를 찾은 후, 이의 역변환된 수치 서열로 치환된다. 이는 아미노산 보존성을 가지면서 워터마크를 삽입하기 위한 DWT 변환 및 역변환 수행 과정이다. 단계별 세부적인 내용은 다음과 같다.

1) 코돈 정수 부호화

DNA 신호 분석을 위한 전처리 과정으로 염기(base), 코돈, 또는 아미노산의 수치 변환이 필요하다. DNA 워터마킹에서는 아미노산 보존성과 변이 강인성을 가지는 워터마크된 DNA를 생성해야 하므로, DWT 및 DCT와 같은 주파수 변환 기반의 워터마킹에서는 위의 염기 수치 사상 기법들로는 적용이 힘들다. 특히 실수, 복소수, 기타 그래픽 표현 방법에 의한 수치 사상은 변환 및 역변환 과정에서 아미노산 보존이 어렵다. 따라서 본 논문에서는 정수 기반 DWT 변환 및 역변환과 보안성, 강인성에 적합한 코돈의 정수 사상을 제안한다.

그림 1(a)의 DNA 유전부호를 살펴보면, 6중 중첩 코돈의 아미노산 R(Arg), S(Ser), L(Leu)에서는 2개의 코돈들이 나머지 코돈과의 거리차가 있어 연속된 정수 할당이 어렵다. 중첩된 코돈들 간의 연속된 정수 할당은 워터마킹 신호 처리를 용이하게 한다. 따라서 제안한 방법에서는 아미노산 R의 {AGG, AGA}, S의 {AGC, AGT}, L의 {TTG, TTT}를 중첩 코돈 중의 하나로 와 같이 임의 치환한다. 그리고 제안한 방법에서는 4개의 염기 b={T,C,A,G}는 Cristea[22~23]의 방법과 유사하게 4개의 정수 ={0,1,2,3} 중 하나씩 임의로 사상한 다음, 각 코돈 $c=(b_1b_2b_3)$을

$$c=4^2\times b_1+4^1\times b_2+4^0\times b_3,\ c\in[0,63] \tag{1}$$

와 같이 정수로 부호화한다. 이 때, 염기 정수 $b_1b_2b_3$들은 코돈 정수 c로부터

$$b_1=\lfloor\frac{n}{4^2}\rfloor,\ b_2=\lfloor\frac{n\%4^2}{4}\rfloor,\ b_3=(n\%4^2)\%4 \tag{2}$$

와 같이 얻을 수 있다. DNA 염기서열에 따라 아미노산의 분포가 다르게 나타나므로, 제안한 방법에서는 아미노산의 히스토그램 분포에 따라 중첩 코돈 단위로 [-32, 31] 범위가 되도록 아미노산 히스토그램 순위 테이블을 설정한 후, 이 테이블에 따라 재배열한다. 예를 들어, 특정 아미노산 A_i의 중첩 코돈들의 정수 부호가 $\Theta_i=\{c_{ij}|A_i=h(c_{ij}),\forall j\in[1,|\Theta_i|]\}$일 때, A_i의 히스토그램 순위는 $rank(A_i)$이라 하고, 이를 x로 놓으면, 이의 역함수 $rank^{-1}(x)=i$는 A_i의 인덱스 i를 나타낸다. 이에 따라 아미노산 히스토그램 순위 테이블에 의하여 재할당된 코돈 정수는

$$\Theta_i'=\{c_{ij}'|A_i=h(c_{ij}),\forall j\in[1,|\Theta_i|]\}, \tag{3}$$
$$\text{where } c_{ij}'=(-32+\sum_{k=1}^{rank(A_i)-1}|\Theta_{rank^{-1}(k)}|)+j$$

와 같다. 여기서 $\Theta_{rank^{-1}(k)}$는 k번째 순위를 가지는 아미노산의 중첩 코돈 정수 부호를 나타낸다. 여기서 순위 테이블은 랜덤하게 생성되며, 워터마크 추출에 필요한 키 정보로 저장된다. 그러므로 아미노산 내의 중첩 코돈 정수 부호는 순위 테이블을 알지 못할 경우 쉽게 예측되지 못할 것이다.

2) 코돈 부호서열의 DWT 변환

DNA 워터마킹에서는 아미노산 보존성을 유지하면서 주파수 변환을 적용하여야 한다. 즉, 워터마크된 주파수 변환 계수에 의하여 역변환된 코돈은 다른 아미노산으로 변경되지 않아야 한다. 따라서 제안한 방법에서는 아미노산 보존성과 함께 변이 강인성을 고려한 정수 Lifting 기반 DWT 변환 계수에 워터마크를 삽입한다.

제안한 방법에서는 그림 3(a)에서와 같이 $N/3$개의 코돈(시작과 끝 코돈 제외)들로 구성된 전체 코돈 부호서열을 $N_B=2^s\ll N/3(s\geq 1)$개 코돈으로 구성된 부-코돈 부호서열 $C_i=\{c_{i1},c_{i2},\cdots,c_{iN_B}\}$으로 분할한다. 이 때, 부-코돈 부호서열의 총 개수는 $floor((N-2)/N_B)$이므로, 이 개수에 맞게 랜덤 워터마크 비트열 $W=\{w_i\in[0,1]|i\in[1,N_W]\}$, $N_W=floor((N-2)/N_B)$이 생성되며, 워터마크 1비트 w_i는 임의의 부-코돈 부호서열 C_i에 차례로 삽입된다.

부-코돈 부호서열 C_i은 정수 lifting 기반 DWT W에 의하여 최대 레벨 $\log_2 N_B$만큼 변환된다. 이때 변환 계수는 저주파 계수 L과, l에서 1번째 레벨까지의 고주파 계수 $H_l,H_{l-1},\cdots,H_1$들로

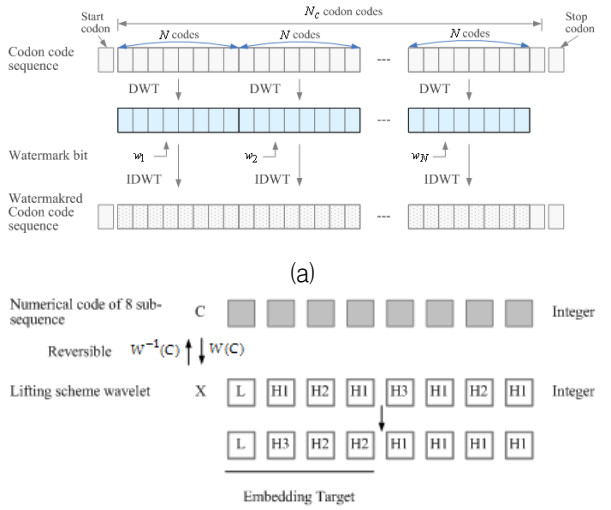


그림 3. (a) 부 코돈 부호서열의 DWT 기반 워터마크 삽입, (b) 크기 8 부-코돈서열의 Lifting 기반 1D-DWT (3레벨)

Fig. 3. (a) Watermark embedding based on DWT of codon sub-sequences and (b) lifting based DWT of codon sub-sequence with 8 codons (3-level).

$$X_i = W(C_i) = \{L, H_l, H_{l-1}, \cdots, H_1\}, \quad (4)$$
$$\text{where } H_k = \{x_{k,1}, x_{k,2}, \cdots x_{k,N/2^k}\}$$

와 같이 분해된다. DWT 기반 영상 워터마킹에서는 강인성과 비가시성을 고려하여 하위레벨보다 상위레벨의 고주파 계수에 워터마크를 삽입한다.

그림 3(b)는 $N_B = 8$일 때 정수 Lifting 기반 DWT의 예를 보여주고 있다. 일반적인 멀티미디어 워터마킹에서는 강인성을 위하여 저주파 계수와 상위레벨의 고주파 계수를 워터마크 삽입 대상으로 선택하며, 연약성을 위하여 하위레벨의 고주파 계수를 삽입 대상으로 선택한다. 이에 따라 제안한 방법에서는 저주파 계수 L과 상위레벨 고주파 계수 H_l, H_{l-1}에 워터마크를 각각 삽입한다. 보안성 향상을 위하여 이들 계수들 중 랜덤하게 선택하여 워터마크가 삽입될 수 있다.

3) 워터마크 삽입

일반적인 DWT 워터마크 삽입 함수인 $x' = (1+\alpha w)x$에서는 아미노산 보존성을 위하여 삽입 강도 α를 조절하여야 한다. 그러나 제한된 중첩 코돈에 의하여 최적의 α를 찾는 것은 매우 어렵다. 따라서 제안한 방법에서는 일반적인 과정과는 반대로, 부-코돈 부호서열 S_i의 아미노산 서열 A_i에 중첩되는 모든 부-코돈 서열의 변환 계수 중에서 워터마크가 삽입되는 가장 적합한 변환 계수를 찾은 후, 이를 역변환하여 워터마크된 부-코돈 부호서열을 생성한다.

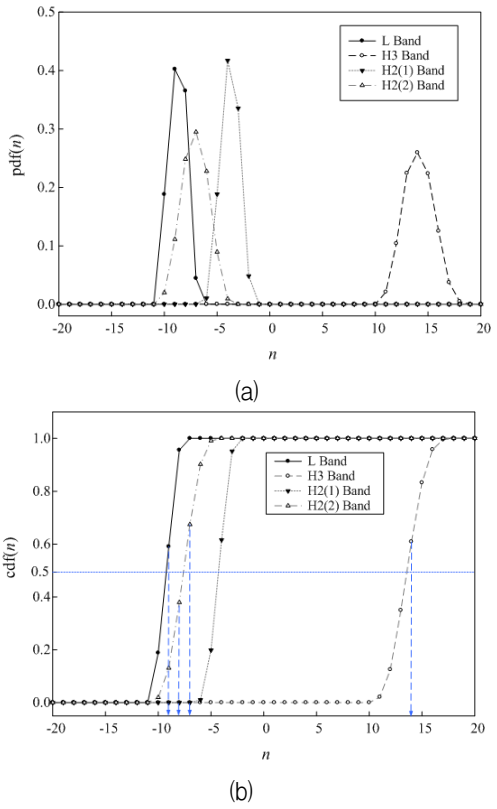


그림 4. Homo Sapiens ANG (NM_001145)에서 N=8일 때, 첫 번째 코돈 제외한 첫 번째 부-코돈 서열인 "GTGATGGGCCTGGGCGTTCTCCTC" (아미노산 VMGLGVLL)의 정수 DWT L 계수, 레벨 2 계수, 레벨 1 계수의 (a) PDF와 (b) CDF

Fig. 4. (a) PDF and (b) CDF of approximation coefficient, detail coefficients of level 1 and level 2 of the first sub-sequence of 8 codons, which sub-sequence is "GTGATGGGCCTGGGCGTTCTCCTC" (Amino acid sequence is "VMGLGVLL"), in Homo Sapiens ANG (NM_001145) sequence.

본 절에서는 하나의 부-코돈 부호서열 S_i의 주파수 계수에 워터마크 1비트 w_i가 삽입되는 방법에 대하여 살펴보기로 한다. 부-코돈서열 C_i에 해당되는 아미노산 서열이 A_i라 할 때, A_i로 번역되는 모든 부-코돈 서열들의 부호 서열 집합 $\boldsymbol{\Omega}_i$(또는 중첩된 부호 서열)과 이들의 DWT 변환 계수 집합 $\mathbf{X}_i$라 하자. 즉, 아미노산 보

존성을 가지는 워터마크된 부-코돈 부호서열 S_i의 DWT 변환 계수는 $\mathbf{X}_i$ 중의 하나가 된다. 여기서 $|\Omega_i|$는 중첩된 부-코돈 서열의 개수(Cardinality)

$$|\Omega_i| = \Pi_{k=1}^{N}|\mathrm{A}_k| \tag{5}$$

으로 각 아미노산의 중첩 코돈 개수 $|\mathrm{A}_k|$들의 곱이다. 수치 사상 과정에서 중첩 코돈 개수가 6이 아미노산들은 4개가 되도록 대체되므로, $|\Omega_i|$의 범위는 $1 \leqq |\Omega_i| \leqq 4^N$이다.

제안한 방법에서는 저주파 계수 또는 하위레벨 고주파 계수 (H_1, H_2)들 중 임의의 $\log_{2^2}N_B$개 계수 $x_{il,k}$ $(l \in [1,2], k \in [1,\log_{2^2}N_B])$를 선택하여 이들 계수에 워터마크 $\log_{2^2}N_B$비트 w_{ik}를 삽입하는 조건을

$$\begin{cases} x'_{il,k} < Th_{il,k}, \text{if } w_{ik} = 0 \\ x'_{il,k} > Th_{il,k}, \text{if } w_{ik} = 1 \end{cases}, \quad \forall k \in [1,\log_{2^2}N_B] \tag{6}$$

와 같이 설정한다. 여기서 $x_{il,k}$에 대한 문턱치 $Th_{il,k}$는 중첩 코돈서열의 변환 계수 집합 $\mathbf{X}_i$에서 l레벨 $(l \in [1,2])$에서 $k(k \in [1,\log_{2^2}N_B])$번째 임의로 선택된 계수 집합 $\boldsymbol{x}_{il,k}$의 CDF, $cdf_{il,k}(n)$가 0.5에 가장 가까운 n으로 $Th_{il,k} = \arg\min_n |cdf_{il,k}(n) - 1/2|$와 같이 정하여 진다. 위 수식에 만족하는 모든 $\log_{2^2}N_B$개의 계수 $x_{il,k}$들을 가지는 변환 계수 그룹 $\mathbf{X}'_i$을

$$\mathbf{X}'_i = \{X'_{i1}, X'_{i2}, \cdots\}, \tag{7}$$
$$\text{where } \{x'_{ij,l,k} | \forall k \in [1,\log_{2^2}N_B]\} \in X'_{ij} = W(C'_{ij})$$

와 같이 모은 다음, 이들의 부-코돈 부호서열 그룹 $\boldsymbol{\Omega}'_i = W^{-1}(\boldsymbol{X}'_i)$을 얻는다. 나머지 과정은 위의 저주파 계수에서 삽입 과정과 동일하다. 즉, 원본 부-코돈 부호서열 C_i는 $\boldsymbol{\Omega}'_i$의 평균 부호서열에 가장 가까운 코돈 부호서열로 대체된다.

그림 4는 ANG 염기서열에서 N=8일 때, 첫 번째 부-코돈 부호 서열 C_i="GTGATGGGCCTGGGCGTTC TCCTC"의 아미노산 A_i="VMGLGVLL"에 중첩되는 모든 부호 서열들의 DWT 변환 계수인 L 계수 $\mathbf{L}_i$, 레벨 2 계수 $\mathbf{H}_{i2}$, 두 개의 레벨 1 계수 $\mathbf{H}_{i1}$들의 pdf와 cdf를 보여주고 있다. 정수 부호 재배열에 의하여 대부분의 계수 분포의 범위가 좁게 나타나지만, CDF에 의한 문턱치에 의하여 동등한 분포 개수를 가지는 두 개의 그룹으로 나누어질 수 있음을 볼 수 있다.

나. 워터마크 추출

의심되는 유전자의 DNA 서열로부터 워터마크 추출은 추출 키인 아미노산 히스토그램 순위 테이블에 의하여 수행된다. 부호 DNA 서열 D^*이 삽입 및 삭제 변이에 의하여 해독틀 변경(Frameshift)이 일어날 경우 워터마크 추출이 매우 어렵다. 따라서 제안한 방법에서는 워터마크 추출하기 전, 의심되는 부호 DNA 서열 D^*을 워터마크된 참조 부호 DNA 서열 D'를 기준으로 Pairwise alignment 알고리즘에 의하여 재배열을 수행한다. 재배열 과정에서 삽입된 부분은 삭제되고, 반대로 삭제된 부분은 D'의 서열로 대체한 후 워터마크 추출 과정이 수행된다.

추출 과정에서는 부호 DNA 서열 D^* 상에 모든 코돈들을 순위 테이블에 의하여 정수로 사상한 다음, 부-코돈 수치 서열 C_i^*의 DWT 변환 계수 $X_i^* = W(C_i^*)$ 상에서 저주파 계수 또는 상위레벨 계수 내에 워터마크 비트 w_i^*를 문턱치에 의하여 추출한다. 즉, 저주파 계수 L_i^*과 하위 레벨 l에 속하는 모든 계수 $H_{il,j}^*$에서 추출된 워터마크 비트의 평균 이진값으로

$$w_i^* = INT\left(\sum_{l}^{|l|}\sum_{j}^{|H_l|} w_{lj}^* / (|H_l| + |l|) + 0.5\right), \tag{8}$$
$$\text{where } w_{ij}^* = \begin{cases} 1, H_{il,j}^* > Th_l \\ 0, H_{il,j}^* \leqq Th_l \end{cases},$$
$$l \in \begin{cases} [\log_2 N_B \log_2 N_B - 1], N_B > 4 \\ \log_2 N_B, \qquad\qquad N_B = 4 \end{cases}$$

와 같이 w_i^*가 추출된다. 각 계수별 문턱치는 앞 절에 설명하였듯이, 워터마크 삽입에서 같이 구하여진다.

Ⅳ. 실험 결과 및 분석

본 실험에서는 In silico 기반으로 표 1에서와 같이 NCBI에서 제공하는 DNA 서열을 사용하여 제안한 방법과 Heider의 DNA-Crypt 기반 워터마킹 방법을 비교 평가하였다. 부-코돈 서열의 길이인 N_B가 커질수록 워터마크에 적합한 서열을 찾는 시간이 매우 길어진다. 따라서 제안한 방법의 실험에서는 $N_B = 4,8$ 길이의 부-코돈 서열들로 분할하였으며, 각 부-코돈 서열별로 워

터마크 비트를 R=3번 반복 삽입하였다. 따라서 N길이의 부호 DNA 서열 상의 워터마크 비트수는 $N_W = floor((N-2)/(RN_B))$. DNA-Crypt 실험에서는 4중 중첩 코돈에 대하여 2비트씩 삽입하였으며, 변이 정정 부호를 위하여 WDH(5)를 사용하였다.

가. 아미노산 보존성 및 용량성

본 실험에서는 표 1의 부호 DNA 서열에 대한 제안한 방법과 Heider 방법의 실험을 1,000번 반복 실험하여 워터마크 전, 후의 염기서열 변화와 아미노산 변화를 살펴보았다. 모든 실험에서 아미노산 서열의 변화는 없음을 확인하였다. 즉, 두 방법 모두 아모노산 서열이 유지되도록 워터마크를 삽입하였으므로, 이는 당연한 결과이다.

표 1. 실험에 사용된 DNA 서열들
Table 1. Tested DNA sequences.

Test Sequence	Gene	CDS 염기수
Bacillus subtilis strain PS beta-glucosidase gene, partial cds	bg1	1407bp
Saccharomyces Cerevisiae S288c Cct2p mRNA, complete cds	CCT2	1584bp
Mycoplasma genitalium genotype 91 adhesin gene, partial cds	mgpB	256bp
Homo sapiens hexosaminidase A (alpha polypeptide), mRNA	HEXA	1590bp
Homo sapiens angiogenin, ribonuclease, RNase A family, 5, transcript variant 1, mRNA	ANG	444bp

DNA 워터마킹은 의도적인 침묵 변이이므로, 표 2에서와 염기서열 변화를 살펴보았다. 유사한 삽입 용량 하에서 제안한 방법에서는 기존 방법에 비하여 3.7배 -5.4배 정도 많은 침묵 변이가 발생하였다. 이는 DWT 변환의 레벨별 받침 영역에 해당되는 모든 코돈들이 워터마크에 의하여 영향을 받기 때문이다. 즉, 제안한 방법은 기존 방법과는 달리 특정 코돈에 하나의 워터마크가 삽입되는 것이 아니고, 여러 개의 코돈들의 조합에 의하여 워터마크가 삽입되므로, 특정 공격에서의 강인성과 보안성에 보다 효과적이다. 침묵 변이가 많이 발생하더라도 아미노산 서열 변화가 없으므로 워터마크 전,후의 유전자는 동일하다. 그림 4는 제안한 H(4,3) 방법 및 DNA-Crypt 알고리즘에 의하여 워터마크된 ANG 서열과 이의 아미노산 서열과 원본 서열과의 Pairwise 정렬된 결과를 보여준다. 이 결과를 살펴보면, 제안한 방법에 의한 염기 서열이 DNA-Crypt 알고리즘보다 많이 변화되나, 아미노산 서열이 유지됨을 볼 수 있다.

Heider의 방법과 제안한 방법과의 삽입 용량 C을 유사하게 실험하기 위하여 본 실험에서는 Heider 방법에서는 WDH의 5번 반복 삽입하였고, 제안한 방법에서는 R=3번 반복 삽입하였다. DNA 서열별 삽입 용량은 표 2에서와 같다.

부-코돈 서열의 총 개수는 $floor((N-2)/N_B)$이고, 이를 M이라 하자. 제안한 방법에서는 N_B=4, R=3일 때,

표 2. DNA의 염기 변화율 및 삽입 용량 (아미노산 변화는 모두 없음, 코돈별 삽입용량 = 워터마크비트수/코돈개수, N_B=부-코돈 서열 개수, R=삽입 반복 횟수)
Table 2. Change rate of bases in watermarked DNA sequences and bit capacity (No amino acids were changed. Bit capacity in codon is [the number of bits / the number of codons], R is the embedding repetition time).

Organism		제안한 방법 (N_B,R)=(4,3)		제안한 방법 (N_B,R)=(8,3)		Heider (WDH(5))
		L대역	H대역	L대역	H대역	
B. Subtilis	염기 변화율	0.580	0.543	0.603	0.580	0.112
	삽입용량	0.081	0.081	0.041	0.122	0.081
S. Cerevisiae	염기 변화율	0.630	0.590	0.615	0.594	0.126
	삽입용량	0.081	0.081	0.040	0.117	0.097
M. Genitalium	염기 변화율	0.480	0.410	0.468	0.621	0.109
	삽입용량	0.071	0.071	0.035	0.106	0.071
Homo Sapiens HEXA	염기 변화율	0.577	0.596	0.601	0.590	0.121
	삽입용량	0.083	0.083	0.042	0.125	0.098
Homo Sapiens ANG	염기 변화율	0.581	0.554	0.628	0.594	0.128
	삽입용량	0.081	0.081	0.041	0.122	0.095

L 대역과 H 대역의 계수가 각각 하나씩 선택되므로, 부-코돈 서열의 총 개수 M의 1/3배에 해당되는 워터마크 비트가 삽입된다. 그러나 N_B=8, R=3일 때, L 대역은 계수 하나가 선택되고, H 대역은 2개 레벨의 3 계수가 선택되므로 L 대역에서는 M의 1/3배, H 대역에서는 M의 워터마크 비트가 삽입된다. 제안한 방법의 삽입 용량 C는 N_B가 4일 때 기존 방법과 유사하며, N_B가 8일 때, L 대역의 용량은 1/2배로 작으나, H 대역은 기존 방법에 비하여 1.2-1.5배 정도 많다.

나. 변이 강인성

변이의 강인성을 평가하기 위하여 본 실험에서는 워터마크된 서열 상에 개시 코돈과 종료 코돈을 제외한 나머지 모든 코돈의 염기들에 대하여 포인트 변이, 삽입 및 삭제 변이를 동시에 발생하였다. 전체 염기 개수에 대한 발생 비율이 r 일 때, 강인성 실험에서는 포인트 변이를 $2r$, 삽입 변이를 r, 삭제 변이를 r로 놓은 후, r을 1-10까지 가변하였다. 예를 들어, r이 5일 때, 5%의 삽입 및 삭제 변이와 10%의 포인트 변이가 동시에 발생된다. 삽입 및 삭제 변이시 해독틀 복원을 위하여 본 실험에서는 워터마크된 서열을 기준으로 변이된 서열을 Smith-Waterman 알고리즘[24]에 의하여 재배열한 다음, 워터마크를 추출하였다. 실험에 사용된 평가 척도로는 추출된 워터마크의 BER(bit error rate)이다.

그림 5. 원본 부호 ANG 시퀀스와의 Pairwise 정렬; 제안한 H대역 (4,3) 방법에 의하여 (a) 워터마크된 ANG 시퀀스 및 (c) 아미노산 시퀀스와 DNA-Crypt에 의하여 (b) 워터마크된 ANG 시퀀스 및 (d) 아미노산 시퀀스

Fig. 5. Pairwsie alignment using original coding ANG sequence; (a) watermarked sequence and (c) amino acid sequence by proposed H band (4,3) scheme, (b) watermarked sequence and (d) amino acid sequence by DNA-Crypt algorithm.

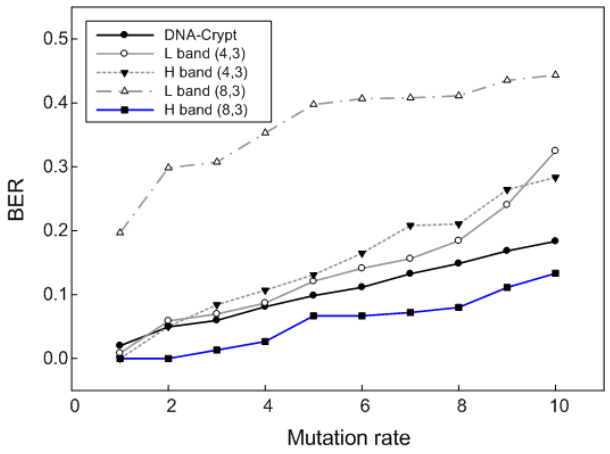


그림 6. 변이된 코돈 서열 상에서 추출된 워터마크의 평균 BER.

Fig. 6. Average BER of watermark extracted in mutated codon sequences.

변이된 DNA 서열 상에서 추출된 워터마크의 BER은 그림 5에서와 같다. 먼저 H 대역 (8,3) 방법은 다른 방법에 비하여 BER이 가장 높게 나타나, 워터마크 삽입 방법으로 적절하지 못하였다. L 대역 (4,3)과 H 대역 (4,3) 방법은 기존 방법에 비하여 변이율이 높을 경우 BER이 높게 나타나, 변이율이 낮을 경우에만 효과적으로 나타났다. 그리고 H 대역 (8,3) 방법은 다른 방법에 비하여 모두 낮은 BER이 나타나, 변이 강인성 성능이 우수하게 나타났다. 즉, 그림 6의 실험 결과로부터 변이율이 낮을 경우, H 대역 (8,3), H 대역 (4,3), L 대역 (4,3), DNA-Crypt, L 대역 (8,3) 순으로 강인성이 나타났으며, 변이율이 높을 경우, H 대역 (8,3), DNA-Crypt, H 대역 (4,3), L 대역 (4,3), L 대역 (8,3) 순으로 강인성이 나타남을 확인할 수 있었다.

DWT 기반 멀티미디어 워터마킹에서는 레벨이 높은 H 대역과 L 대역이 다른 대역에서보다 우수한 강인성을 가진다. 이와 반대로 DNA 워터마킹에서는 웨이블릿 필터의 반침영역에 속하는 코돈의 변화에 민감하게 나타나므로, 낮은 레벨의 H 대역에 워터마크를 삽입하는 것이 변이 강인성이 우수하게 나타났다.

다. 보안성

워터마크의 보안성은 삽입 기법 및 서열이 공개되더라도 워터마크 추출이 어려워야 한다. 그러나 기존 방법은 워터마크된 위치를 쉽게 예측할 수 있어, 워터마크 제거가 용이하다. 제안한 방법에서는 아미노산 히스토그램의 랜덤 순위표가 삽입 키로 사용된다. 이 순위표의 개수는 20!=2.4329e18이며, 특정 삽입에 사용된 순위표가 없을 경우, 코돈 수치 및 DWT 변환 계수를 찾기가 어렵다. 보안성을 더욱 향상시키기 위하여, N_B 부-코돈 서열을 랜덤하게 선택하거나, 또는 DWT 변환 계수들 중 랜덤하게 선택된 계수에 워터마크를 삽입하는 등 다양한 방법에 의하여 제시될 수 있다. 또한 기존 생물학적 DNA 암호화에서와 같이 PCR 프라이머 쌍 또는 특정 프로브를 암호키로 삽입된 서열 위치를 암호함으로써 생물학적 DNA 워터마킹으로 확장될 수 있다.

V. 결 론

본 논문에서는 강인성 및 보안성을 가지는 DWT 기반 부호 DNA 서열 워터마킹 기법을 제안하였으며, DWT 변환 및 역변화 과정에서 아미노산이 보존되도록 워터마크가 삽입될 수 있음을 나타내었다. 제안한 방법에서는 부-코돈 서열을 아미노산 히스토그램 순위 테이블에 의한 정수 변환한 다음, 부-코돈 부호 서열의 DWT 변환 계수 중 L 대역과 저레벨의 H 대역에 워터마크를 삽입한다. 이 때, 워터마크된 DWT 계수로 인한 아미노산 변경 방지를 위하여, 부-코돈 서열의 동위 코돈 서열들의 DWT 변환 계수 집합 중에 워터마크 삽입에 가장 적합한 동위 코돈 서열을 찾은 후, 이를 부-코돈 서열과 치환한다. 본 실험에서는 4개 코돈의 부-코돈 서열의 L 대역과 레벨 1의 H 대역 계수에 워터마크를 3번 반복하여 삽입하였고 (L 대역 (4,3), H 대역 (4,3)), 8개 코돈의 부-코돈 서열의 L 대역 계수와 레벨 1,2의 H 대역 계수들 중 랜덤하게 선택된 계수에 워터마크를 3번 반복하여 삽입하였다(H 대역 (4,3), H 대역 (8,3)). 실험 결과로부터 H 대역 (8,3) 방법만이 DNA-Crypt 방법에 비하여 강인성과 용량성이 우수하였으며, 워터마크 키인 아미노산 히스토그램 순위 테이블에 의한 보안성과 아미노산 보존성을 만족하였다.

참 고 문 헌

[1] T. Kazuo, O. Akimitsu, and S. Isao, “Public-key systems using DNA as a one-way function for key distribution,” *BioSystems*, vol. 81, no. 1, pp. 25-29, 2005.

[2] M. Yamamoto, S. Kashiwamura, A. Ohuchi and M. Furukawa, “Large-scale DNA memory based on the nested PCR,” *Natural Computing*, vol. 7, no. 3, pp. 335-346, 2008.
[3] A. Gehani, T. LaBean and J. Reif, “DNA-based Cryptography,” Aspects of Molecular Computing, *Lecture Notes in Computer Science*, 2004, vol. 2950/2004, pp. 34-50, 2004.
[4] C. T. Clelland, V. Risca, C. Bancroft, “Hiding messages in DNA microdots,” *Nature*, vol. 399, pp. 533-534, June 1999.
[5] A. Leier, C. Richter, W. Banzhaf, and H. Rauhe, “Cryptography with DNA binary strands,” *Biosystems*, vol. 57, Issue 1, pp. 13-22, June 2000.
[6] V. I. Risca, “DNA-based steganography,” *Cryptologia,* vol. 25, no. 1, pp. 37-49, 2001.
[7] M. Arita and Y. Ohashi, “Secret signatures inside genomic DNA,” *Biotechnol. Prog*. vol. 20, pp. 1605-1607, 2004.
[8] G.C. Smith, C.C. Fiddes, J.P. Hawkins, and J.P. Cox, “Some possible codes for encrypting data in DNA,” *Biotechnology letters*, vol. 25, no. 14. pp. 1125-1130, July 2003.
[9] N. Yachie, K. Sekiyama, J. Sugahara, Y. Ohashi, and M. Tomita, “Alignment-based approach for durable data storage into living organisms,” *Biotechnol. Prog.* vol. 23, pp. 501-505, April 2007.
[10] N. Yachie, Y. Ohashi, and M. Tomita, “Stabilizing synthetic data in the DNA of living organisms,” *Systems and Synthetic Biology,* vol. 2, no. 1-2, pp. 19-25, June 2008.
[11] D. Heider and A. Barnekow, “DNA watermarks in non-coding regulatory sequences,” *BMC Bioinformatics*, vol. 2, no. 125, 2009.
[12] D. Heider and A. Barnekow, “DNA-based watermarks using the DNA-Crypt algorithm,” *BMC Bioinformatics*, vol. 8, no. 176, May 2007.
[13] D. Heider and A. Barnekow, “DNA Watermarks - A proof of concept,” *BMC Bioinformatics*, vol. 9, no. 40, April 2008.
[14] D. Heider, D. Kessler, and A. Barnekow, “Watermarking sexually reproducing diploid organisms,” *Bioinformatics*, vol. 24, no. 17, pp. 1961-1962, 2008.
[15] 이석환, 권성근, 권기룡, “부호 영역 DNA 시퀀스 기반 강인한 DNA 워터마킹,” 대한전자공학회논문지, 제49권 CI편 제2호, pp. 123-132, 2012년 3월.
[16] 김정연, 남제호, “DCT 압축영역에서의 DC 영상 기반 다해상도 워터마킹 기법,” 대한전자공학회, 전자공학회논문지-SP, 제45권 제4호, pp. 1-9, 2008년 7월.
[17] 박혜정, 최준림, “H.264/AVC 비디오 보호를 위한 비가시적 워터마킹의 설계 및 검증,” 대한전자공학회, 전자공학회논문지-SD, 제45권 제6호, pp. 74-79, 2008년 6월
[18] 이석환, 권기룡, “기하학적 구조 및 위치 보간기를 이용한 3D 애니메이션 워터마킹,” 대한전자공학회, 전자공학회논문지-CI, 제43권 제6호, pp. 71-82, 2006년 11월.
[19] 이석환, 권성근, 권기룡, “볼록 집합 투영 기법을 이용한 3D 메쉬 워터마킹,” 대한전자공학회, 전자공학회논문지-CI, 제43권 제2호, pp. 81-92, 2006년 3월.
[20] Genetic code, http://en.wikipedia.org/wiki/Genetic_code
[21] T.A. Brown, Genomes 3, Garland Science, 2006.
[22] P.D. Cristea, “Conversion of nucleotides sequences into genomic signals,” *Journal of Cellular and Molecular Medicine*, vol. 6, issue. 2, pp. 279-303, April 2002.
[23] P.D Cristea, “Genetic signals an emerging concept,” *Proc. of IWSSIP*, pp. 17-22, 2011.
[24] T. Smith and M. Waterman, “Identification of common molecular subsequences,” *Journal of Molecular Biology*, vol. 147, pp. 195-197, 1981.

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 전자컴퓨터정보통신공학부 교수
<주관심분야 : 멀티미디어 정보보호, 멀티미디어 통신 및 신호처리>

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 조교수
<주관심분야: 멀티미디어 암호, 모바일 방송, 워터마킹>