논문 2012-49CI-2-16

부호 영역 DNA 시퀀스 기반 강인한 DNA 워터마킹

(Robust DNA Watermarking based on Coding DNA Sequence)

이 석 환*, 권 성 근**, 권 기 룡***

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

요 약

본 논문에서는 DNA 시퀀스의 불법 복제 및 변이 방지와 개인 정보 침해 방지, 또는 인증을 위한 DNA 워터마킹에 대하여 논의하며, 변이에 강인하고 아미노산 보존성을 가지는 부호영역 DNA 시퀀스 기반 DNA 워터마킹 기법을 제안한다. 제안한 DNA 워터마킹은 부호 영역의 코돈 서열에서 정규 특이점에 해당되는 코돈들을 삽입 대상으로 선택되며, 워터마크된 코돈이 원본 코돈과 동일한 아미노산으로 번역되도록 워터마크가 삽입된다. DNA 염기 서열은 4개의 문자 {A,G,C,T}로 (RNA은 {A,C,G,U}) 구성된 문자열이다. 제안한 방법에서는 워터마킹 신호처리에 적합한 코돈 부호 테이블을 설계하였으며, 이 테이블에 따라 코돈 서열들을 정수열로 변환한 다음 원형 각도 형태의 실수열로 재변환한다. 여기서 코돈은 3개의 염기들로 구성되며, 64개의 코돈들은 20개의 아미노산으로 번역된다. 선택된 코돈들은 아미노산 보존성을 가지는 원형 각도 실수 범위 내에서 인접 코돈과의 원형 거리차 기준으로 워터마크에 따라 변경된다. HEXA와 ANG 시퀀스를 이용한 *in silico* 실험을 통하여 제안한 방법이 기존 방법에 비하여 아미노산 보존성을 가지면서 침묵 변이와 미스센스 변이에 보다 강인함을 확인하였다.

Abstract

This paper discuss about DNA watermarking using coding DNA sequence (CDS) for the authentication, the privacy protection, or the prevention of illegal copy and mutation of DNA sequence and propose a DNA watermarking scheme with the mutation robustness and the animo acid preservation. The proposed scheme selects a number of codons at the regular singularity in coding regions for the embedding target and embeds the watermark for watermarked codons and original codons to be transcribed to the same amino acids. DNA base sequence is the string of 4 characters, {A,G,C,T} ({A,G,C,U} in RNA). We design the codon coding table suitable to watermarking signal processing and transform the codon sequence to integer numerical sequence by this table and re-transform this sequence to floating numerical sequence of circular angle. A codon consists of a consecutive of three bases and 64 codons are transcribed to one from 20 amino acids. We substitute the angle of selected codon to one among the angle range with the same animo acid, which is determined by the watermark bit and the angle difference of adjacent codons. From in silico experiment by using HEXA and ANG sequences, we verified that the proposed scheme is more robust to silent and missense mutations than the conventional scheme and preserve the amino acids of the watermarked codons.

Keywords: DNA 워터마킹(DNA watermarking), DNA 정보보호 (DNA security), 코돈 부호 테이블(Codon coding table), 변이 공격(Mutation attack), 부호 DNA 시퀀스 (Coding DNA Sequence)

* 정회원, 동명대학교 정보보호학과
(Dept. of Information security, Tongmyong University)
** 정회원-교신저자, 부경대학교 전자컴퓨터정보통신공학부
(Div. of Electronics, Computer & Telecommunication, Pukyong National University)
*** 정회원, 경일대학교 전자공학과
(Dept. of electronics engineering, KyungIl University)
※ 이 논문은 2011년도 정부(교육과학기술부)의 재원으로 한국연구재단의 기초연구사업 지원을 받아 수행된 것임 (KRF-2011-0023118, KRF-2011-0010902)
접수일자: 2011년11월22일, 수정완료일: 2012년3월6일

Ⅰ. 서 론

DNA 내에 포함된 유전 코드(genetic code)는 '인간의 일기'라고 불리어지는 심오한 개인 정보로, 타인에게 노출될 경우 프라이버시를 포함한 인권 침해가 심각할 것이다. 따라서 유전 정보의 수집절차 적법성과 공정성 그리고 제도적인 안정망이 구축되어야 하며, 제한된 범위에서의 유전자 정보 활용 및 대외기밀 유지 등을 위한 대책이 마련되어져야 한다[1]. HGI(human genome information) 사용에 대한 윤리적인 법 또는 가이드라인 이외에도 HGI의 불법 복제/도용 방지 및 인증을 위한 기술이 필요하다. 또한 1g의 DNA는 108 Tera byte를 저장할 수 있을 만큼 DNA는 대용량의 정보를 저장할 수 있는 새로운 생체 정보 매개체로 인식되어짐에 따라 DNA 암호화 시스템에 대한 필요성이 제기되고 있다. 이와 같은 필요성에 의하여 최근 유전자재조합유기체(GMO, genetically modified organism)의 암호 및 정보 은닉을 위하여 DNA/RNA 시퀀스인 A, G, T(U), C 문자열에 대한 암호화[2~3], 스테가노그라픽 및 워터마킹 기법[4~11]들이 제안되어지고 있다. DNA 스테가노그라피 및 워터마킹 기법들은 실험 환경에 따라 생물학적 실험인 *in vivo/vitro* (within the living)와 컴퓨터 내 실험인 *in silico*으로 구분되어진다. 대부분 바이오 전공가들은 *in vivo* 기반의 DNA 정보 은닉 기법[2~3, 5~7, 9~10]을 제안하였고, 신호 처리 전공가들은 *in silico* 기반의 기법[8, 11]을 제안하고, 실험으로 증명하였다.

유기체(Organism)의 게놈(Genome)은 한 개체 유전자의 총 염기서열로, 단백질로 번역(translation)되는 부호 DNA (coding DNA, cDNA)과 그렇지 않은 비부호 DNA(non-coding DNA, ncDNA)으로 구분되어진다. 따라서 cDNA[8~11] 또는 ncDNA[5~7]에 따라 정보를 은닉하는 방법이 달라져야 한다. ncDNA은 단백질로 번역되지 않고 Gene을 포함하지 않으므로, 정보가 은닉된 ncDNA가 임의로 추가되거나 정보에 따라 부분 ncDNA의 임의 치환이 가능하다. 최근 연구가들은 ncDNA의 많은 부분들이 유전 체계의 중요한 기능을 제어하는 데 포함되므로 실제적인 'Junk'가 아니라고 밝혀지고 있다. 그럼에도 불구하고, ncDNA은 유기체의 단백질 프로파일을 변경하지 않는다고 가정되어지고 있다. 위의 가정에 따르면, ncDNA은 어떤 정보가 삽입되더라도 유기체의 속성에는 변함이 없으므로, 비밀 메시지 전송을 위한 스테가노그라피에 적용이 가능하나 강인성을 요구하는 워터마킹에는 적합하지 못하다. cDNA는 유전자 부호(Genetic code)에 의하여 단백질로 번역되므로, 유기체의 단백질 프로파일을 유지하면서 정보가 은닉되어야 한다. 이는 cDNA 정보 은닉의 필수 제한 조건으로, 아미노산 (Amino acids) 보존성 또는 코돈 동의성 (Codon equivalence)이라 한다. 본 논문에서는 아미노산 보존성이라 하기로 한다. 일반 영상 및 비디오, 3D 워터마킹[14~17]과는 달리 변이 공격에 강인한 DNA 워터마킹 설계시 가장 어려운 부분이 아미노산 보존성이다.

기존 DNA 워터마킹 기법들을 살펴보면, 단순한 치환 기법 또는 유전 부호 기반의 심볼 및 코돈들의 비트 할당에 의한 워터마크 삽입과 *in vivo* 기반의 워터마크된 셀 구현에 초점을 맞추고 있다. 그러므로 신호처리 관점에 강인성, 가시성 및 용량성과 같은 평가 분석과 이들 조건을 만족하는 DNA 워터마킹 기법이 필요하다.

본 논문에서는 아미노산 보존성을 유지하면서 변이에 강인한 cDNA 시퀀스 워터마킹 기법을 제안하며, 이를 *in silico* 기반으로 평가 분석하였다. 제안한 방법에서는 기존 유전자 부호와는 달리 순회 원형 부호로 코돈들을 배열한 다음, 차례로 정수 할당한다. 그리고 코돈별 각도 특이점에 따라 연속된 3개 코돈들의 집합을 선택한 후, 연속된 3개 코돈들 간의 순회 원형 부호 내 각도 차이에 따라 워터마크 비트를 삽입한다. 이 때 워터마크 비트에 따라 아미노산 보존성을 유지하면서 3개 코돈들이 변경된다. 제안한 순회 원형 부호는 DNA 심볼의 정수 변환 및 역변환이 용이하고, 임의 위치에서의 심볼 에러시 예측이 가능하고, 동일한 아미노산을 생성하는 코돈들이 인접한 정수로 할당되도록 설계되어진다. *in silico* 기반의 평가 실험에서 제안한 방법이 아미노산 보존성을 유지하며, 기존 cDNA 방법에 비하여 치환, 삽입, 및 삭제 변이에 강인함을 확인하였다.

본 논문의 구성은 다음과 같다. Ⅱ장에서는 DNA 염기서열 구조, 유전자 부호, 및 DNA 워터마킹 기법들에 대하여 분석하고, Ⅲ장에서는 제안한 DNA 워터마킹 기법에 대하여 자세히 살펴보기로 한다. Ⅳ장에서는 *in silico* 기반 다양한 DNA들에 대한 평가 분석을 살펴보고 마지막으로 Ⅴ에서는 본 논문의 결론을 맺는다.

Ⅱ. DNA 워터마킹

1. DNA 유전 부호

유전정보를 담고 있는 DNA와 RNA는 길게 이어져 있는 뉴클레오티드(nucleotide)로 이루어져 있으며, 각 뉴클레오티드는 인(phosphate), 당(sugar), 염기(base)로 구성되어 있다. 그 중 DNA의 염기는 A (Adenine), G(Guanine), C(Cytosine), T(Thymine)의 네 종류가 있고 RNA의 경우 U(Uracil)가 T를 대체하고 있다. 코돈(codon)이라고 하는 세 개의 염기서열은 유전 부호(Genetic code)에 의하여 하나의 아미노산이 결정된다. 즉, DNA 유전 부호는 그림 1에서와 같이 단백질 합성 과정에서 DNA나 RNA의 염기서열을 아미노산 서열로 바꿔주는 규칙으로, 트리플릿 코드(triplet code)라고도 한다. 염기는 A,G,C,T의 4가지이므로, 총 43=64개의 코돈이 존재하며, 이들 코돈들은 20개의 아미노산 중 하나를 지정하거나 전사(Transcription)가 끝나는 종료코돈(stop codon)을 지정하게 된다[18~19].

각 아미노산에 따라 코돈 길이가 1에서 6까지이다. 여기서 부호 영역을 시작하는 시작 코돈은 Met (ATG)(진핵 및 원핵 세포)이고, 부호 영역의 끝을 나타내는 종료코돈은 Stp (TAA, TAG, TGA)이다. cDNA 내에 워터마크 삽입시 시작 및 종료 코돈을 제외한 나머지 코돈들에서 아미노산 보존성을 가지는 코돈들 간의 치환에 의하여 가능하다.

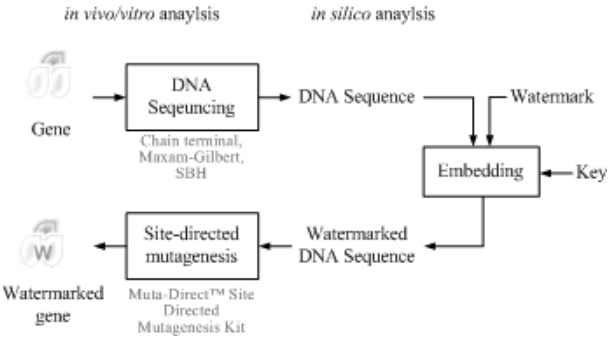


그림 1. 워터마크된 유전자 생성 과정
Fig. 1. The generation process of watermarked gene.

2. DNA 워터마킹

in vivo 기반 DNA 워터마킹은 그림 1에서와 같이 Sanger-Coulson 기법(사슬 종결법), Maxam-Gilbert, 또는 SBH(Sequencing by Hybridization)의 자동염기서열법에 의하여 유전자로부터 DNA 시퀀스를 획득한 다음, 워터마크에 의해 변이된 DNA 시퀀스에 따라 Site-directed mutagenesis와 같은 방법에 의하여 워터마크된 유전자를 획득한다. *in silico* 기반 DNA 워터마킹은 DNA 시퀀스와 워터마크로부터 워터마크된 시퀀스를 획득하여 이를 평가 분석한다.

영상/비디오[14~15] 및 3D 워터마킹[16~17]에서와 같이 DNA 시퀀스 기반 워터마킹에서도 다음과 같은 주요 요구 조건들이 있다.

1) 아미노산 보존성 : 원본 DNA와 워터마크된 DNA와의 아미노산 시퀀스는 동일하여야 한다.

2) 강인성 : 워터마크는 DNA 삽입, 복제, 삭제 및 치환 변이 등에 강인하여야 하며, DNA 정보 훼손에도 워터마크가 추출되어야 한다.

3) 메시지 충실도 : DNA 시퀀스와 워터마크 정보 간의 상호 간섭이 없어야 하며, DNA 표현형이 유지되어야 한다.

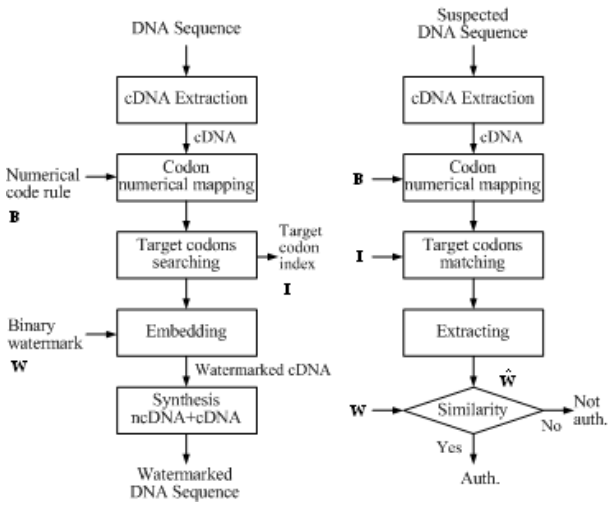
4) 검출 오류 허용치 : 전송 지연에 의한 유기체 변이 또는 생물체 실험 (PCR, 유기체 합성 등) 환경 하에서 워터마크 검출 오류가 발생될 수 있다. 검출 오류 허용치 내에 워터마크가 복원될 수 있어야 한다.

5) 전달 및 해석 용이 : 워터마크된 DNA 시퀀스가 명확하여야 하며, 이로 이루어진 유기체 내에 워터마크 검출 및 전달이 용이하여야 한다.

6) 보안성 : 제3자에 의한 워터마크 추출이 확률적으로 불가능하도록 DNA 워터마크의 보안성이 확보되어야 한다.

7) 용량성 : 충분한 정보가 삽입될 수 있도록 워터마크의 용량이 확보되어야 한다.

본 논문에서는 위의 요구 조건들 중 아미노산 보존성, 강인성, 보안성을 가지는 *in silico* 기반 DNA 워터마킹 기법을 제안하며, 나머지 조건들에 대하여 간략히 살펴보기로 한다.



(a) (b)

그림 2. 제안한 DNA 시퀀스 워터마크 (a) 삽입 및 (b) 추출 과정

Fig. 2. The proposed process for (a) embedding and (b) extracting of watermark.

Ⅲ. DNA 워터마킹

본 논문에서는 DNA 시퀀스의 워터마킹 기법을 제안하며, 다음과 같은 주요 특징을 가진다. 1) cDNA 시퀀스에 이진 워터마크가 삽입된다. ncDNA 시퀀스는 유전 체계의 주요한 기능을 조절하는데 포함되므로, ncDNA 보다 cDNA가 워터마크에 삽입 대상에 적합하다. 2) 워터마크된 코돈들이 아미노산 보존성을 가진다. 3) 포인트 변이 및 삽입/삭제 변이에 강인하다.

제안한 워터마크 삽입 과정은 그림 2(a)에서와 같이 원형 형태의 코돈 실수 변환, DWT 기반 특이 검출에 의한 삽입 대상 코돈 선택 및 코돈별 워터마크 삽입 및 ncDNA와 워터마크된 cDNA와의 결합에 의한 워터마크된 DNA 시퀀스 생성의 단계로 구성된다. 삽입 대상 코돈의 인덱스는 삽입 추출시 필요한 키로 저장된다. 이때 키는 cDNA 시퀀스 변경에 영향을 주지 않기 위하여 DNA 시퀀스와 별도로 저장되거나, ncDNA의 더미 시퀀스에 치환되어 저장되어진다. 워터마크는 그림 2(b)에서와 같이 삽입 과정과 유사한 과정으로 추출된다.

본 논문에서 사용되는 주요 기호의 정의는 다음과 같다. 뉴클레오티드의 염기는 b={G,A,C,T}이고, 세 염기들의 서열인 코돈은 $c=b_1b_2b_3$이다. 코돈 c의 아미노산은 $S=f(c)$으로 f|C→S는 표 1에서와 같이 코돈에서 아미노산의 번역이다. |S|는 아미노산 S으로 번역되는 코돈들의 개수를 나타낸다.

1. 코돈 정수 변환

염기와 코돈 시퀀스들은 G, A, C, T들의 문자열과 같으므로, 이진, 정수 또는 실수 형태의 워터마크 삽입을 위하여 염기 및 코돈들을 정수 또는 실수 형태로 변환하여야 한다. 제안한 방법에서는 표 1에서와 같이 6비트의 정수로 할당한 후, 이를 원형 형태의 각도로 변환한다. 먼저 4개의 염기들은 G=0, A=1, C=2, T=3으로 놓은 다음, 각 코돈 $c=b_1b_2b_3$은 다항식 형태에 의하여 6비트 정수 n로

표 1. 16×4 행렬 형태의 코돈 부호 테이블

Table 1. Codon code table of 16x4 matrix type.

S	c	n	code	S	c	n	code	S	c	n	code	S	c	n	code
G	GGG	0	000	R	AGG	16	100	R	CGG	32	200	W	TGG	48	300
	GGA	1	001		AGA	17	101		CGA	33	201	Stop	TGA	49	301
	GGC	2	002	S	AGC	18	102		CGC	34	202	C	TGC	50	302
	GGT	3	003		AGT	19	103		CGT	35	203		TGT	51	303
E	GAG	4	010	K	AAG	20	110	Q	CAG	36	210	Stop	TAG	52	310
	GAA	5	011		AAA	21	111		CAA	37	211		TAA	53	311
D	GAC	6	012	N	AAC	22	112	H	CAC	38	212	Y	TAC	54	312
	GAT	7	013		AAT	23	113		CAT	39	213		TAT	55	313
A	GCG	8	020	T	ACG	24	120	P	CCG	40	220	S	TCG	56	320
	GCA	9	021		ACA	25	121		CCA	41	221		TCA	57	321
	GCC	10	022		ACC	26	122		CCC	42	222		TCC	58	322
	GCT	11	023		ACT	27	123		CCT	43	223		TCT	59	323
V	GTG	12	030	M	ATG	28	130	L	CTG	44	230	L	TTG	60	330
	GTA	13	031	I	ATA	29	131		CTA	45	231		TTA	61	331
	GTC	14	032		ATC	30	132		CTC	46	232	F	TTC	62	332
	GTT	15	033		ATT	31	133		CTT	47	233		TTT	63	333

$$n = 4^2 \times b_1 + 4^1 \times b_2 + 4^0 \times b_3 \quad (1)$$

와 같이 할당한다. 그리고 정수 n에 의하여 코돈을 원형 각도 $g(c)$로 $g(c) = \frac{2n\pi}{64}$ or $\frac{2n\pi}{64} - 2\pi$와 같이 변환한다. 여기서 그림 3에서와 같이 두 각도 중 이전 코돈 c_{-1}의 $g(c_{-1})$와의 차이가 작은 것으로 할당된다.

$$g(c) = \begin{cases} \frac{2n\pi}{64}, & \text{if} \left|\frac{2n\pi}{64} - g(c_{-1})\right| < \left|\frac{2n\pi}{64} - 2\pi - g(c_{-1})\right| \\ \frac{2n\pi}{64} - 2\pi, & \text{if} \left|\frac{2n\pi}{64} - g(c_{-1})\right| > \left|\frac{2n\pi}{64} - 2\pi - g(c_{-1})\right| \end{cases} \quad (2)$$

원형 각도 $g(c)$로부터 코돈 c를 얻기 위한 역 변환에서는 먼저 정수 n를

$$n = \begin{cases} \frac{64}{2\pi} g(c), & \text{if } g(c) > 0 \\ \frac{64}{2\pi}(2\pi + g(c)), & \text{otherwise} \end{cases} \quad (3)$$

구한 다음, 코돈의 세 염기를 $b_1 = \lfloor n/4^2 \rfloor$, $b_2 = \lfloor (n\%4^2)/4 \rfloor$, $b_3 = (n\%4^2)\%4$와 같이 추출한다. 그리고 $b_1 b_2 b_3$로부터 코돈 c가 구하여진다.

2. 워터마크 삽입

가. 삽입 대상 코돈 탐색

제안한 방법에서는 워터마크 1비트 당, $w_i = \{0,1\}$, 3개의 연속 코돈 그룹 (ternion) $C_i = \{c_{k-1}, c_k, c_{k+1}\}$에 삽입한다. 워터마크가 삽입되는 코돈들은 임의로 선택될 수 있으나, 제안한 방법에서는 코돈 원형 각도 $g(c)$들의 특이점에 해당되는 코돈인 임계화된 원형 각도 시퀀스로부터 국부 최대치 LM가

$$LM(c_k) = g(c_k), \text{ if } (l_k > th \text{ and } r_k > th) \text{ and } (\forall S(c_{k-1}), S(c_k), S(c_{k+1}) \neq 'W \text{ and } g'(c_{k-1}) \neq LM) \quad (4)$$
$$\text{where } th = \alpha \times m$$

와 같이 구하여진다. 국부 최대치 LM는 인접 코돈의 원형 각도가 국부 최대치가 아니고, $|S|>1$ 인 코돈들 중 좌,우 인접 코돈 간의 각도 차이가 문턱치 th보다 큰 $g'(c)$로 정의된다. $|S|=1$인 아미노산 $S='W'$의 코돈 'TGG'는 삽입 대상에서 제외된다. 여기서 문턱치 th의 m은 인접한 코돈 간의 평균 각도 차이이고, α는 워터마크 비트수를 조절하는 계수이다. 즉, α에 의하여 문턱치를 조절함으로써 워터마크 비트수를 결정한다. 제안한 방법에서는 LM 위치의 코돈과 인접 두 코돈들을 묶어 삽입 대상 코든 그룹으로 선택한다.

$$C_i = \{c_{k-1}, c_k, c_{k+1}\} \quad (5)$$
$$\text{where } LM(c_{k-1}), LM(c_{k+1}) = 0, \; LM(c_k) = g'(c_k)$$

cDNA 시퀀스 내에 삽입되는 워터마크 비트수와 LM의 개수 $|LM|$와 동일하다.

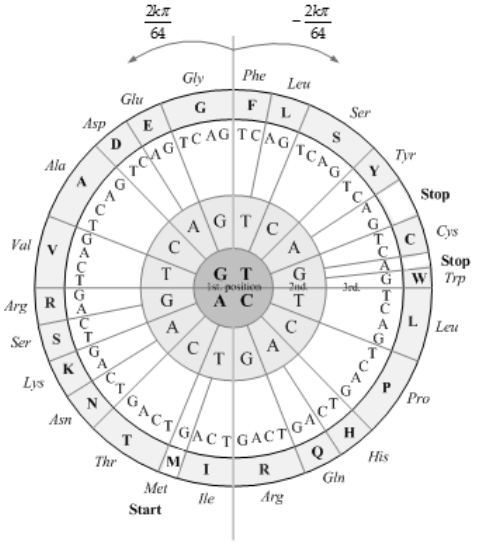


그림 3. 코돈별 순회 원형 코드
Fig. 3. Recursive circle code for codon.

나. 워터마크 비트 삽입

제안한 방법에서는 워터마크 비트 w_i를 삽입 대상 코돈 그룹 $C_i = \{c_{k-1}, c_k, c_{k+1}\}$의 코돈들 내에 각각 삽입한다. 중간 코돈에서는 아미노산이 가지는 최대 및 최소 원형 각도에 따라 w_i가 삽입된다. 그리고 좌,우 코돈들은 중간 코돈을 기준점으로 시계 방향 또는 반시계 방향으로의 최대 및 최소 거리를 가지는 원형 각도에 따라 w_i가 삽입된다.

먼저 제안한 방법에서는 세 코돈들의 아미노산 $f(c_{k-1}) = S_{k-1}, f(c_k) = S_k, f(c_{k+1}) = S_{k+1}$을 얻는 다음, 세 아미노산들 내의 코돈 원형 각도 $g(S_{k-1}), g(S_k), g(S_{k+1})$들을 얻는다. 중간 코돈 c_k의 아미노산 S_k의 평균 원형 각도들은 $R_k = \frac{1}{|S_k|} \sum_{j=1}^{|S_k|} g(c_{k,j})$와 같다. 그림 3의 순회 원형 코드 상에서 R 기준으로 시계방향으로 S_{k-1} 내의 코돈 원형 각도들과의 거리 $d\cup$와 반시계방향으로 S_{k+1} 내의 코돈 원형 각도들과의 거리 $d\cup$는

$$d\circlearrowleft(R_k,g(c_{k-1,j}))=|g(c_{k-1,j})-R_k|,\ \forall\, j\in[1,|S_{k-1}|],\quad (6)$$
$$d\circlearrowright(R_k,g(c_{k+1,j}))=2\pi-|g(c_{k+1,j})-R_k|,\ \forall\, j\in[1,|S_{k+1}|]$$

와 같이 정의된다. S_{k-1} 내에 시계방향으로 R_k과의 최대 및 최소 거리를 가지는 코돈 $c_{k-1,\max}$, $c_{k-1,\min}$들과 S_{k+1} 내에 반시계방향으로 R_k과의 최대 및 최소 거리를 가지는 코돈 $c_{k+1,\max}$, $c_{k+1,\min}$들을 구한다. S_k 내의 최대 및 최소 원형 각도는 $c_{k,\min}=g(c_{k,1}), c_{k,\max}=g(c_{k,|S_k|})$와 같다. 워터마크 비트 w_i는 $C_i=\{c_{k-1},c_k,c_{k+1}\}$에 각 코돈들이

$$C'_i=\begin{cases}\{c_{k-1,\min},c_{k,\min},c_{k+1,\min}\}, & \text{if } w_i=0\\ \{c_{k-1,\max},c_{k,\max},c_{k+1,\max}\}, & \text{if } w_i=1\end{cases} \quad (7)$$

와 같이 변경됨으로써 삽입된다. 이와 같은 방법에 의하여 선택된 모든 삽입 대상 코돈 집합들에 워터마크 비트가 각각 삽입된다.

3. 워터마크 추출

전송 또는 의심스러운 DNA 시퀀스로부터의 워터마크 추출 과정은 그림 2(b)에서와 같이 삽입 과정과 유사한다. 추출 과정에서는 cDNA 시퀀스들의 모든 코돈들을 원형 각도로 변환한 후, DWT hard 임계화된 원형 각도 시퀀스들의 국부 최대치에 해당되는 코돈들을 탐색한다. 그리고 삽입 과정에서 저장된 삽입 대상 코돈의 인덱스 $\mathbf{I}$와 탐색된 코돈의 인덱스 $\hat{\mathbf{I}}$와의 동기화한 다음, 동기화된 인덱스 $\mathbf{I}$로부터 삽입 대상 코돈 그룹들을 구한다. 삽입 및 삭제 변이에 의하여 시퀀스들이 쉬프트되므로, $\mathbf{I}$와 $\hat{\mathbf{I}}$를 차례로 매칭하면서 불일치한 코돈부터 한 코돈씩 쉬프트함으로써 동기화가 수행된다.

각 코돈 그룹 $C_i=\{c'_{k-1},c'_k,c'_{k+1}\}$에서 코돈들의 아미노산 $f(c'_{k-1})=S'_{k-1}$, $f(c'_k)=S'_k$, $f(c'_{k+1})=S'_{k+1}$과 원형 각도 $g(S'_{k-1})$, $g(S'_k)$, $g(S'_{k+1})$와 이들의 평균 각도 $\bar{g}(S'_{k-1})$, $\bar{g}(S'_k)=R_k$, $\bar{g}(S'_{k+1})$들을 각각 구한 다음, 중간 코돈의 평균 각도 R_k를 기준으로 시계방향으로 $g(c'_{k-1})$와의 거리 $d\circlearrowleft(R,g(c'_{k-1}))$와 반시계방향으로 $g(c'_{k+1})$와의 거리 $d\circlearrowleft(R_k,g(c'_{k+1}))$를 구한다. 그리고 좌,우 및 중간 코돈 내의 삽입 비트들을 R_k와의 비교함으로써

$$w'_{k-1}=\begin{cases}0, & \text{if } d\circlearrowleft(R_k,g(c'_{k-1}))<d\circlearrowleft(R_k,\bar{g}(S'_{k-1}))\\ 1, & \text{otherwise}\end{cases} \quad (8)$$

$$w'_k=\begin{cases}0, & \text{if } g(c'_k)<R_k\\ 1, & \text{otherwise}\end{cases} \quad (9)$$

$$w'_{k+1}=\begin{cases}0, & \text{if } d\circlearrowleft(R_k,g(c'_{k+1}))<d\circlearrowleft(R_k,\bar{g}(S'_{k+1}))\\ 1, & \text{otherwise}\end{cases} \quad (10)$$

구한다. 위의 세 비트들의 다수결 원칙에 의하여 코돈 그룹 내에 삽입된 워터마크 비트 w'_i가 추출된다.

이상의 추출 과정에 의하여 모든 삽입 대상 코돈 그룹으로부터 워터마크 $\hat{\mathbf{W}}$이 모두 추출된다. 원본 워터마크 $\mathbf{W}$와 추출된 워터마크 $\hat{\mathbf{W}}$와의 유사도 검출 또는 비트 에러율에 의하여 DNA 시퀀스의 저작권 위배 유무가 결정된다.

Ⅳ. 실험 결과

본 실험에서는 표 2에서와 같이 NCBI에서 제공하는 Homo Sapiens 시퀀스의 CDS(Coding sequence)를 사용하여 제안한 방법과 Heider[7, 10]의 DNA-Crypt 기반 워터마킹 방법을 비교 평가하였다. 제안한 방법에서는 국부 최대치 탐색을 위한 문턱치 변수 α를 0.1으로 놓은 후, 삽입 대상 코돈 그룹을 결정하였다. Heider의 DNA-Crypt 기법에서는 오류 정정을 위한 8/4 Hamming 부호화가 사용되었으며, 코돈 개수가 4 이상인 아미노산 {G,A,V,T,R,P,L,S}으로 번역되는 모든 코돈들에 워터마크가 삽입되었다. 모든 실험들은 Matlab Bioinformatics toolbox 3 기반으로 수행되었다.

그림 4는 원본 ANG 시퀀스와 제안한 방법 및 Heider 방법에 의하여 워터마크된 ANG 시퀀스들을 보여주고 있다. 이 그림을 살펴보면, 제안한 방법과 Heider 방법에 의하여 워터마크된 ANG 염기 서열들은 원본 염기 서열에 비하여 21%와 11% 정도 차이가 나지만, 두 방법 모두 ANG 아미노산 서열들은 원본 서열

표 2. 실험에 사용된 DNA 시퀀스
Table 2. Tested DNA sequences.

GenBank accession number	NM_000520	NM_001145
Gene	HEXA	ANG
Organism	Homo sapiens	
총 염기수	2437bp	1222bp
CDS 구간과 염기수	[208-1797], 1590bp	[601-1044], 444bp

그림 4. ANG(NM_001145)에 대하여 제안한 방법의 (a) 워터마크된 염기 서열과 (b) 아미노산 서열 및 Heider 방법의 (c) 워터마크된 염기 서열과 (d) 아미노산 서열

Fig. 4. ANG(NM_001145) sequence; (a) nucleotide base sequence and (b) amino acid sequence watermarked by the proposed method, (c) nucleotide base sequence and (d) amino acid sequence watermarked by Heider's method.

과 일치함을 볼 수 있다. 따라서 두 방법 모두 아미노사 보존성이 유지됨을 알 수 있다. 또한 DNA 시퀀스와 워터마크 간의 상호 간섭이 없도록 워터마크가 생성되므로 메시지의 충실도가 보장된다.

1. 변이 강인성

강인성 평가 실험에서는 인위적인 침묵 변이와 미스센스 변이를 수행하였다[18~20].

1) 침묵 변이

침묵 변이 실험에서는 CDS 내의 모든 코돈들 중에 일정 비율 γ 만큼 임의로 선택한 후, 이들 코돈들을 동

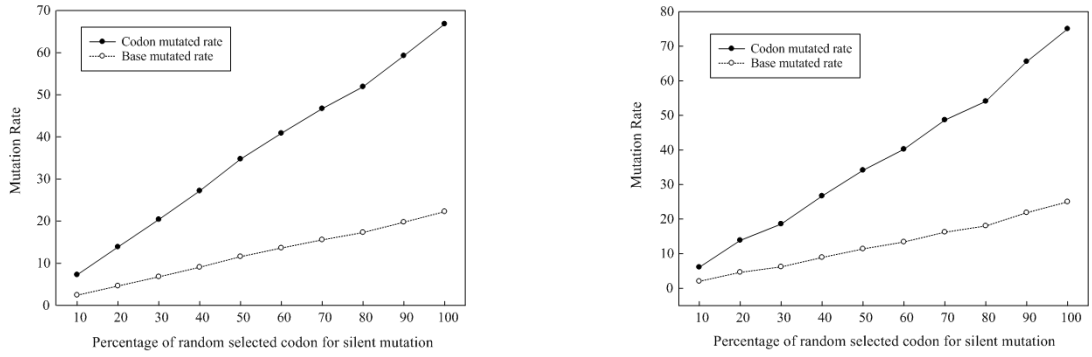


그림 5. 침묵 변이된 (a) HEXA와 (b) ANG의 코돈 및 염기 변이율
Fig. 5. Mutation ratio of codon and base of silent mutated (a) HEXA와 (b) ANG.

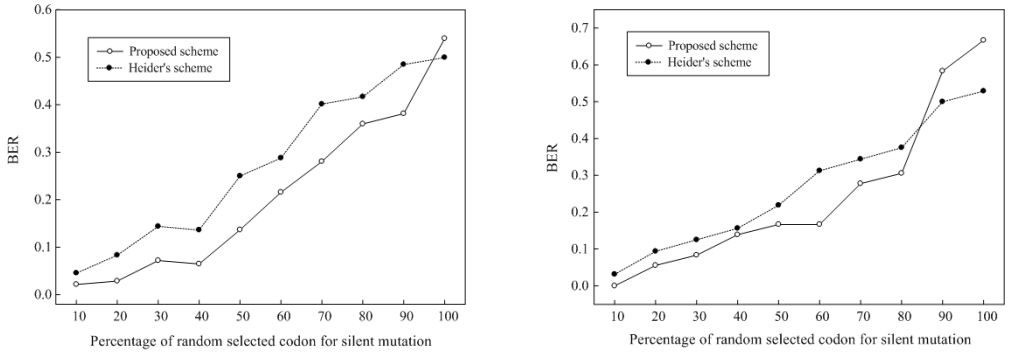


그림 6. 제안한 방법 및 Heider 방법에 의하여 침묵 변이된 (a) HEXA와 (b) ANG에서 추출된 워터마크의 BER
Fig. 6. BERs of extracted watermarks from silent mutated (a) HEXA와 (b) ANG by proposed scheme and Heider's scheme.

일 아미노산을 가지는 코돈들 중에 하나로 대체하였다. 예를 들어, 'GTT' 코돈일 경우, 표 1에서와 같이 아미노산 'V'로 번역되며, 'V'로 번역되는 코돈들은 {GTG, GTA, GTC, GTT}으로 코돈 길이는 4이다. 이들 코돈들 중 균일 분포의 랜덤 함수에 의하여 4개 중 하나로 선택되어 이로 대체된다. 따라서 침묵변이로 선택된 코돈들 중 변이되지 않는 코돈들이 발생되며, 이 확률은 코돈 길이의 역수에 해당된다. 실험에서는 γ를 10%-100%로 가변하면서 침묵 변이를 수행하였다.

그림 5(a), 5(b)는 HEXA, ANG 시퀀스에 대하여 선택 코돈 비율 γ에 따른 코돈 변이율과 염기 변이율을 나타내고 있다. 침묵 변이에서 코돈 내의 한 염기가 변경되므로 염기 변이율은 코돈 변이율의 1/3에 해당된다. 그림에서 살펴보면, γ이 100%일 때, 코돈 변이율은 66-75%이고, γ이 50%일 때, 코돈 변이율은 약 34%이고, 염기 변이율은 약 11%이다.

침묵 변이에 대한 실험 결과인 그림 6(a), 6(b)를 살펴보면, γ이 10%-90%일 때, 제안한 방법이 Heider 방법에 비하여 1.15-2.89배 정도 BER이 낮게 나타났다. 예를 들어 γ이 50%일 때 제안한 방법의 BER은 0.1367-0.1667이고, Heider 방법의 BER은 0.21-0.25이다. 따라서 본 실험을 통하여 제안한 방법이 기존 방법에 비하여 침묵 변이에 강인함을 알 수 있었다.

2) 미스센스 변이

미스센스 변이 실험에서는 침묵 변이 실험에서와 같이 일정 비율 γ 만큼 임의로 선택한 후, 이들 코돈들을 64개 코돈들 중 임의의 코돈으로 대체하였다. 선택된 코돈이 바뀌지 않을 확률은 약 1/64이며, 코돈 내에 세 개의 염기가 모두 바뀌거나 또는 하나, 두 개의 염기가 바뀔 수 있다. 따라서 코돈 변이율은 γ와 거의 동일하다. 실험에서는 γ을 10%-50%으로 가변하면서 미스센스 변이를 수행하였다. 미스센스 변이 실험 결과인 그림 7(a), 7(b)를 살펴보면, γ이 10%-30%일 때, 제안한

방법과 Heider 방법의 BER이 0.1이하로 낮게 나타났다. 그러나 γ가 40%-50%으로 높을 때 제안한 방법의 BER은 0.1111-0.1583으로 다소 낮으나 Heider 방법의 BER은 0.1563-0.2576으로 높게 나타났다. 따라서 제안한 방법이 기존 방법에 비하여 미스센스 변이에 강인함을 알 수 있었다.

2 삽입 용량

CDS 내에 삽입되는 워터마크의 용량은 아미노산 보존성과 변이에 대한 강인성을 고려하여 적절하게 선택되어야 한다. 제안한 방법에서는 삽입 대상 코돈 그룹인 3개 코돈 당 1비트 삽입되므로, CDS 내에 시작 코돈과 종료 코돈을 제외한 나머지 코돈들을 3개 코돈으로 그룹화하여 1비트 삽입하면 최대 삽입 용량은 C_{max} =1/3[bit/codon]이다. 제안한 방법은 보안성 향상과 코돈 서열의 특이점에 해당되는 코돈 선택을 위하여 DWT Hard 임계화된 코돈 서열에서 국부 최대치를 사용한다. 국부 최대치를 결정하는 문턱치는 α 변수에 의하여 결정되며, 국부 최대치 개수 $|LM|$에 의하여 워터마크 삽입 용량이 결정된다. 따라서 제안한 방법의 워터마크 삽입 용량 C는

$$C=\frac{1}{3}\times\frac{|LM|}{(|C|-2)/3}=\frac{|LM|}{|C|-2}\ \text{[bit/codon]} \qquad (11)$$

와 같다. 여기서 $|C|$-2는 CDS 내에 시작 및 종결 코돈을 제외한 코돈 개수를 나타낸다. 삽입 대상 코돈 그룹 내에 하나의 코돈에 워터마크 1비트씩 삽입이 가능하며, 이 때 워터마크 삽입 용량 C는 $\frac{|LM|}{(|C|-2)/3}$이다. 제안한 방법에서는 워터마크 추출 위하여 삽입 코돈 인덱스의 정보가 부가적으로 필요하다. 그러나 이 정보는 DNA 시퀀스와 별도로 저장되거나, ncDNA 시퀀스의 더미(dummy) 정보에 저장되어 전송되어지므로 cNDA 시퀀스에는 전혀 영향을 주지 않는다. 따라서 부가적인 정보에 대한 용량은 삽입 용량 계산에서 제외되었다.

Heider의 방법은 워터마크 비트를 8/4 Hamming 부호화에 의하여 부호한 다음, 코돈 길이 $|S|$가 4 이상인 아미노산의 코돈들에 각각 2비트씩 차례로 삽입한다. 그러므로, 워터마크 삽입 용량 C는

$$C=\frac{1}{2}\times\frac{2|Z|}{|C|-2}=\frac{|Z|}{|C|-2}\ \text{[bit/codon]} \qquad (12)$$

와 같다. 여기서 $|Z|$는 아미노산 {G,A,V,T,R,P,L,S}으로 번역되는 코돈의 개수를 나타낸다. HEXA 및 ANG 시퀀스에 대하여 제안한 방법은 Heider 방법에 비하여 삽입 용량이 약 0.52-0.56배 정도 낮으나, 삽입에 사용되는 코돈 개수는 1.68-1.95배 정도 높게 나타났다. 즉, 제안한 방법은 Heider 방법에 비하여 삽입 용량이 다소 낮으나 워터마크 된 코돈 개수가 높으므로, 강인성이 보다 우수하게 나타남을 확인하였다.

Ⅴ. 결 론

부호영역 DNA 워터마킹의 가장 주요한 조건으로 아미노산 보존성과 변이에 대한 강인성이 있다. 본 논문에서는 위의 두 조건을 만족하는 DNA 워터마킹 기법을 제안하였다. 제안한 방법에서는 코돈 부호 테이블 설계, 부호영역 코돈 서열의 정수 변환과 원형 각도의 실수 변환, DWT Hard 임계화된 코돈들의 국부 최대치 기반 정규 특이점 코돈 탐색, 아미노산 보존성 규칙에 따른 워터마크 삽입의 단계로 이루어져 있다. 코돈 부호 테이블은 DNA의 워터마킹 신호 처리에 적합하게 설계되어졌으며, 워터마크는 인접 코돈의 원형 각도 간에 따라 삽입됨으로써 침묵 변이 및 미스센스 변이에 강인성을 가지게 되었다. 제안한 방법의 성능 평가를 위한 실험에서는 Homo sapiens의 HEXA와 ANG 시퀀스를 이용하여 침묵 변이 및 미스센스 변이에 대한 강인성, 아미노산 보존성 및 워터마크의 용량성을 평가하였다. 실험 결과로부터 제안한 방법이 기존 방법에 비하여 우수한 변이 강인성을 가지며, 두 방법 모두 아미노산 보존성을 가짐을 확인하였다. 그러나 제안한 방법은 워터마크 용량이 다소 낮으나 기존 방법과는 달리 용량 조절이 가능하며, 보안성을 가질 수 있음을 확인하였다.

제안한 방법에서는 워터마크 추출시 삽입 대상 코돈의 인덱스 정보가 필요하다. 즉, DNA 시퀀스 이외에 부가 정보가 저장되어야 할 공간이 필요하다. 부가 정보는 cDNA 시퀀스의 손상없이 DNA 시퀀스와 별도로 저장되거나, 또는 ncDNA 시퀀스에 치환에 의하여 저장될 수 있다. 향후 연구에서는 보다 정확한 부가 정보 저장과 전송에 대하여 논의하고자 하며, 워터마크의 보안성에 대한 정량적인 평가를 수행하고자 한다.

참 고 문 헌

[1] P. Sankar, “Genetic Privacy,” *Annual Review of Medicine*, vol. 54, pp. 393-407, Feb. 2003.

[2] C. T. Clelland, V. Risca, C. Bancroft, “Hiding messages in DNA microdots,” *Nature*, vol. 399, pp. 533-534, June 1999.

[3] A. Leier, C. Richter, W. Banzhaf, and H. Rauhe, “Cryptography with DNA binary strands,” *Biosystems*, vol. 57, Issue 1, pp. 13-22, June 2000.

[4] B. Anam, K. Sakib, M. A. Hossain, and K. Dahal, “Review on the Advancements of DNA Cryptography,” *4th International Conference on Software, Knowledge, Information Management and Applications*, Aug. 2010.

[5] 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.

[6] D. Heider and A. Barnekow, “DNA watermarks in non-coding regulatory sequences,” *BMC Bioinformatics*, vol. 2, no. 125, 2009.

[7] D. Heider and A. Barnekow, “DNA-based watermarks using the DNA-Crypt algorithm,” *BMC Bioinformatics*, vol. 8, no. 176, May 2007.

[8] B. Shimanovsky, J. Feng, and M. Potkonjak, “Hiding data in DNA,” *Procs. of the 5th Intl. Workshop in Information Hiding*, pp. 373-386, October 2002.

[9] M. Arita and Y. Ohashi, “Secret Signatures Inside Genomic DNA,” *Biotechnology Prog.*, vol. 20, pp. 1605-1607, 2004.

[10] D. Heider and A. Barnekow, “DNA Watermarks - A proof of concept,” *BMC Bioinformatics*, vol. 9, no. 40, April 2008.

[13] J. Shuhong and R. Goutte, “Code for encryption hiding data into genomic DNA of living organisms,” *9th International Conference on Signal Processing (ICSP)*, pp. 2166-2169, Oct. 2008.

[14] 김정연, 남제호, “DCT 압축영역에서의 DC 영상 기반 다해상도 워터마킹 기법,” 대한전자공학회, 전자공학회논문지-SP, 제45권 제4호, pp. 1-9, 2008년 7월.

[15] 박혜정, 최준림, “H.264/AVC 비디오 보호를 위한 비가시적 워터마킹의 설계 및 검증,” 대한전자공학회, 전자공학회논문지-SD, 제45권 제6호, pp. 74-79, 2008년 6월

[16] 이석환, 권기룡, “기하학적 구조 및 위치 보간기를 이용한 3D 애니메이션 워터마킹,” 대한전자공학회, 전자공학회논문지-CI, 제43권 제6호, pp. 71-82, 2006년 11월.

[17] 이석환, 권성근, 권기룡, “볼록 집합 투영 기법을 이용한 3D 메쉬 워터마킹,” 대한전자공학회, 전자공학회논문지-CI, 제43권 제2호, pp. 81-92, 2006년 3월.

[18] T.A. Brown, *Genomes 3*, Garland Science, 2006.

[19] D. Anastassiou, “Genomic Signal Processing,” *IEEE Signal Processing Magazine*, pp. 8-20, July 2001.

[20] R.C. Deonier, S. Tavar´e, S, and M.S. Waterman, *Computational Genome Analysis: An Introduction*, Springer, 2005.

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 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
관심분야: 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 컴퓨터전자공학부 부교수
2006년～현재 부경대학교 전자컴퓨터정보통신공학부 교수
<주관심분야 : 멀티미디어 정보보호, 멀티미디어 통신 및 신호처리>