Digital Signal Processing 25 (2014) 173–189

Contents lists available at ScienceDirect

Digital Signal Processing

www.elsevier.com/locate/dsp

DNA sequence watermarking based on random circular angle

Suk-Hwan Lee *

Department of Information Security, Tongmyong University, 535, Yongdang-Dong, Namgu, Busan, 608-711, Republic of Korea

ARTICLE INFO

Article history:
Available online 4 December 2013

Keywords:
DNA watermarking
Copyright protection
Codon random mapping table
Amino acid residue conservation
Mutation resistance

ABSTRACT

This paper discusses DNA watermarking for copyright protection and authentication of a DNA sequence. We then propose a DNA watermarking method that confers mutation resistance, amino acid residue conservation, and watermark security. Our method allocates codons to random circular angles using a random mapping table and selects a number of codons for embedding targets using the Lipschitz regularity that is measured from the evolution across scales of local modulus maxima of codon circular angles. We then embed the watermark into random circular angles of codons without changing the amino acid residue. The length and location of target codons depend on the random mapping table and the singularity of detection of Lipschitz regularity. This table is used as the watermark key and can be applied to any codon sequence regardless of sequence length. Without knowledge of this table, it is very difficult to detect the length and location of sequences for extracting the watermark. From experimental results on the suitability of similar watermark capacities, we verified that our method has a lower bit-rate error for point mutations compared with previous methods. Further, we established that the entropies of the random mapping table and the location of target codons are high, indicating that the watermark is secure.

1. Introduction

The genetic code contains profound personal information. It may be considered as a personal diary, and its unauthorized disclosure may be considered as a grave invasion of privacy and a violation of human rights. Legal measures have been established to ensure the safety and security of procedures for collecting human genetic information (HGI) [1–3]. There are laws of ethics or guidelines for HGI use, but security techniques for preventing illegal copying and piracy of HGI are urgently required. DNA is considered as a new biometric medium for storing extraordinarily large amounts of data. Thus, DNA storage demands that DNA security techniques are addressed. Recent research on DNA security includes DNA cryptography [4–7], steganography [8–16], and watermarking [17–25] using DNA sequences with a character stream of A, G, T (or U), and C. These studies [4–25] were validated by *in vivo*, *in vitro*, or *in silico* experiments.

DNA cryptography [4–7] is a technique for biological encryption and decryption based on the polymerase chain reaction (PCR) or DNA chips and has been recognized as a new biological encryption technique for potential widespread use in the future. However, it has not yet begun to replace the conventional encryption algorithms because of the difficulty in its implementation. DNA steganography [8–16] is the technique for hiding messages in DNA sequences and is useful for DNA signature/identification and DNA storage of vast quantities of information. However, the purposes of DNA cryptography and steganography are not to recover DNA sequences or messages under changing experimental conditions or from mutations, and they are therefore not suitable as applications for copyright protection.

DNA watermarking [17–25] is a technique for protecting the information within a DNA sequence. DNA-based watermarks can be applied to copyright-protect DNA sequences as well as for discriminating between wild-type and artificial genomes [25]. Recently, J. Craig Venter's research team inserted a watermark at intergenic sites known to tolerate transposon insertions to identify the genome of *Mycoplasma genitalium* JCVI-1.0 as synthetic [26,27]. Jupiter et al. [28,29] presented the strategy of watermark implementation for tracking select or infectious agents. They suggested five features for implementation strategy as follows: message fidelity, error tolerance, easy interpretation, uniqueness, and resistance. They compared and analyzed the features of DNA and multimedia watermarking. Multimedia watermarking schemes for audio, image, video, and 3D model are mostly processed in the frequency domain of the discrete cosine transform (DCT) [30,31], discrete wavelet transform (DWT) [32–34], and scale-invariant feature transform (SIFT) [35], as well as the geometric domain [36, 37]. However, DNA-based watermark must be embedded without changing the function of coding regions, which represents the main difference between DNA and multimedia watermarking.

* Fax: +82 51 628 1129.
E-mail addresses: skylee@tu.ac.kr, sukhwanlee@gmail.com.

1051-2004/$ – see front matter
http://dx.doi.org/10.1016/j.dsp.2013.11.010

The genome contains all of an organism's hereditary information, and it includes coding sequences (genes), which are translated into polypeptide chains (proteins), representing, for example, approximately 1.5% of the human genome. Sequences that do not encode proteins may be transcribed into non-coding RNAs such as micro-RNAs, which may code for RNAs that regulate gene function. The genome also includes pseudogenes, which are mutated remnants of genes that are unable to encode a functional product. The majority of the human genome comprises non-coding repeated sequences. DNA steganography or watermarking methods have been designed differently depending on whether the information is embedded in non-coding [17–20] or coding DNA [21–24].

Non-coding DNA-based methods typically assume that non-coding DNA does not change the phenotype of an organism. Based on this assumption, any pirate could substitute an arbitrary sequence or dummy sequence for a non-coding DNA sequence while including embedded sequences without changing phenotype. However, many types of non-coding DNA sequences have known biological functions, including the transcriptional and translational regulation of protein-coding sequences. Non-transcribed DNA sequences may contribute to chromosomal properties or possess functions that are yet to be discovered. Therefore, it is not yet been appropriate to manipulate non-coding DNA for steganography and watermarking.

Coding DNA is translated to a polypeptide chain, so the watermark embedded in coding DNA sequences should preserve the protein profile of an organism. This is a prerequisite and a limitation of coding DNA-based methods. We refer to this limit here as amino acid residue (amino acid) conservation. Amino acid conservation is problematic when designing DNA watermarking, in contrast to image and video watermarking.

Most conventional DNA watermarking methods focus on the embedding process through simple substitutions or bit allocations to a base or codon as determined by the genetic code, and focus has been placed on watermarked gene in *in vivo* experiments. Therefore, it is necessary to analyze the resistance to phenotypic change, amino acid sequence conservation, as well as security for signal processing when designing a DNA watermarking method that satisfies the above requirements.

Here we present a coding DNA watermarking method for copyright protection of a DNA sequence that provides mutation resistance, watermark security, and amino acid sequence conservation. We analyze the performance of coding DNA watermarking using *in silico* experiments. The main features of our method are as follows: First, we map codons to numerical values of random circular angles using a random mapping table for security and ease of signal processing. The random mapping table for 64 codons, which includes start and stop codons, is used as the watermark key. Coding sequences of various lengths can be mapped to random circular angles using any random mapping table. Second, we select a number of target codons for embedding using the Lipschitz regularity of local modulus maxima at multi-resolution scales. The local modulus maxima of random circular angles depend on the random mapping table. The length and location of target codons depend on Lipschitz values of local modulus maxima. It is very difficult to detect locations of embedded codons without the knowledge of random mapping table. Third, we embed repeatedly a binary watermark into random circular angles of codons to confer mutation resistance. An angle of a center codon and the distance between two angles of neighboring codons are changed by a bit of a watermark without changing the encoded amino acid. Finally, random circular angles are based on circular coding, which makes the numerical transformation of DNA symbols easier and allows estimation of symbol errors in arbitrary positions. Moreover, it allows the allocation of synonymous codons to neighboring numerical values.

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

Fig. 1. DNA genetic code.

The performance of our *in silico* experiments verified that our method ensures amino acid sequence conservation and is more resist to point mutations compared with DNA-Crypt watermarking [21] and the Liss method [24]. We investigated the capacity based on the analysis by Balado et al. [38–40]. We computed the entropy of the random mapping table and random positions of codons for analyzing security and confirmed that the entropies were high.

This paper is organized as follows: In Section 2, we explain the structure of DNA sequences and the genetic code and analyze conventional watermarking methods. We present the proposed DNA watermarking method in Section 3 before analyzing its performance using *in silico* experiments in Section 4. Finally, are conclusions are presented in Section 5.

2. The genetic code and DNA watermarking

In this section, we examine the genetic code [41–43] and then analyze the requirements of DNA watermarking for copyright protection of a DNA sequence.

2.1. The genetic code

DNA is a long polymer that carries genetic information in simple units called nucleotides, the backbones of which are composed of carbohydrate and inorganic phosphate groups joined by phosphodiester bonds. The nucleotide bases are the purines A (adenine) and G (guanine), and the pyrimidines C (cytosine) and T (thymine). Uracil (U) occurs in RNA instead of thymine. Three nucleotides form a codon, the basic unit of the genetic code that specifies one amino acid. Fig. 1 shows that the genetic code represents a set of rules, which allows the information encoded in DNA to be transcribed into messenger RNAs (mRNAs) that are translated into polypeptide chains composed of amino acids. This code is called the triplet code. There are $4^3 = 64$ different codon combinations, because there are four distinct nucleotide bases. All 64 codons encode one of 20 amino acids or translational stop signals.

Table 1 shows the correspondence of codons to amino acids (Fig. 1). In this table, the cardinality of synonymous codons in each amino acid is from 1 to 6. 'ATG' encoding methionine ('Met') is the most frequently used start codon, whereas 'TAA', 'TAG', and 'TGA' for 'Stp' are stop codon marking the end of coding regions. Except for start and stop codons and the codon for the tryptophan (W), all codons in coding DNA sequences can be used for watermark embedding, and they can be substituted by a watermark while conserving the amino acid. However, all codons including the excepted codons are allocated to numerical values and circular angles because of computational tractability of numerical mapping.