Hindawi
Security and Communication Networks
Volume 2018, Article ID 3530969, 13 pages
https://doi.org/10.1155/2018/3530969

Research Article

Reversible Data Hiding for DNA Sequence Using Multilevel Histogram Shifting

Suk-Hwan Lee

Department of Information Security, Tongmyong University, Busan 608-711, Republic of Korea

Correspondence should be addressed to Suk-Hwan Lee; skylee@tu.ac.kr

Received 11 October 2017; Revised 3 January 2018; Accepted 4 February 2018; Published 13 March 2018

Academic Editor: Emanuele Maiorana

A large number of studies have examined DNA storage to achieve information hiding in DNA sequences with DNA computing technology. However, most data hiding methods are irreversible in that the original DNA sequence cannot be recovered from the watermarked DNA sequence. This study presents reversible data hiding methods based on multilevel histogram shifting to prevent biological mutations, preserve sequence length, increase watermark capacity, and facilitate blind detection/recovery. The main features of our method are as follows. First, we encode a sequence of nucleotide bases with four-character symbols into integer values using the numeric order. Second, we embed multiple bits in each integer value by multilevel histogram shifting of noncircular type (NHS) and circular type (CHS). Third, we prevent the generation of false start/stop codons by verifying whether a start/stop codon is included in an integer value or between adjacent integer values. The results of our experiments confirmed that the NHS- and CHS-based methods have higher watermark capacities than conventional methods in terms of supplementary data used for decoding. Moreover, unlike conventional methods, our methods do not generate false start/stop codons.

1. Introduction

DNA has the potential for high-capacity and long-term storage, which is of considerable interest to researchers in a wide range of applications related to biology and information technology [1–3]. Rapid progress in synthesizing defined strings of DNA and sequencing the data saved in DNA has enhanced the feasibility of DNA storage. In addition to DNA storage [1–3], DNA steganography for secret communication or encryption using noncoding DNA sequences [4–6] has been widely studied for decades. Recent years have witnessed the use of coding DNA sequences to investigate DNA watermarking for tracking parent genes in offspring or genetically modified organisms and for protecting DNA copyright [7–14]. In silico and in vitro/in vivo tracking of offspring by watermarking has been easily realized in bacteria and other genetically modified genomes. However, watermarking of sexually reproducing organisms remains an open issue because recombination events can destroy the watermark. Heider et al. [8] identified a coupled Y-chromosomal/mitochondrial DNA watermarking procedure as the most appropriate DNA watermarking procedure for diploid organisms by using population predictions and statistical analyses, but there remains a lack of experimental validation in this regard. Balado [15, 16] modeled the Shannon capacity of DNA data embedding under mutations for irreversible DNA watermarking.

A common consideration in DNA storage, steganography, and watermarking is how to embed the external data in a DNA sequence to preserve the biological function for various purposes. Furthermore, reversible data hiding for DNA storage, steganography, or watermarking is necessary to recover an original DNA sequence without loss of information. However, most methods [1–14] are irreversible. In addition to facilitating original DNA recovery, reversible DNA data hiding can prevent DNA forgery and mutations from external data, and the mutation process can be analyzed using the iterative process of embedding, detecting, and recovering the information. Reversible image data hiding (or watermarking) has been investigated in many studies using difference expansion [21], prediction error expansion [22–24], histogram shifting [25, 26], lossless compression

[27], and quantization index modulation [28], among other methods. In addition, the performance analysis of relevant methods has been reported [29], and it has been shown that expansion-based methods are more effective than other methods in terms of capacity, imperceptibility, and computation. However, reversible DNA data hiding has not been investigated as extensively as reversible image data hiding or irreversible DNA data hiding for DNA storage, steganography, and watermarking because of the small quantities of four-character-symbol nucleotide bases and their limited reversibility.

With regard to reversible DNA data hiding, Chen [19] encoded a sequence of nucleotide bases in noncoding DNA into decimal values and embedded a watermark in coded values using a lossless compression and difference expansion (DE) algorithm described by Tian [21]. Further, Huang et al. [20] used a histogram with low modification rates. Lee and Kwon presented consecutive DE multiple bit embedding (CDE-MBE) [17] and least-squares-based prediction error expansion (LS-PE) [18] of neighbor code values of noncoding DNA sequences while preventing a false start codon. They reported that LS-PE has 0.36 bits per nucleotide base (bpn) more than CDE-MBE [18]. Other methods [30–33] use substitution by the complementary rule of base pairs with a reference DNA sequence. Although these methods cannot change the length of a DNA sequence, they can introduce false start codons [19, 20, 30–33], require a reference sequence to detect and recover nonblind sequences, [31–33], or have extremely low watermark capacity [20].

In this study, we examine reversible DNA data hiding methods using histogram shifting (HS). In addition, we aim to not only prevent false start/stop codons but also achieve blind detection/recovery and high watermark capacity. First, we encode nucleotide bases with four-character symbols into $2n$-bit values as a unit of n nucleotide bases. Thus, we can easily handle the nucleotide bases. Next, we embed multiple bits in each integer value using multilevel histogram shifting of noncircular type (NHS) and circular type (CHS). Unlike an image, a sequence of integer values of a DNA sequence does not have a regular histogram distribution; hence, multilevel histogram shifting is possible by numeric coding. These methods provide a high watermark capacity and facilitate blind detection and recovery. Finally, we verify whether a false start/stop codon is generated in an integer value and between consecutive integer values.

Through experiments, we evaluate the capacity efficiency of watermark bpn versus extra data bpn and the occurrence of false start/stop codons for our NHS- and CHS-based methods as well as for the methods described by Chen [19], Huang et al. [20], and Lee et al. (LS-PE) [18]. Extra data capacity is required for detecting the watermark and recovering the original sequence. The capacity efficiency is the number of watermark bits that can be embedded per bit of extra data. The experimental results show that the CHS-based method, NHS-based method, LS-PE-based method, Chen's method, and Huang's method have watermark capacities and capacity efficiencies of 0.584 bpn and 1.818, 0.409 bpn and 1.239, 0.419 bpn and 0.234, 0.108 bpn and 0.495, and 0.027 bpn and 0.131, respectively. In addition, we find that our methods do not introduce false start/stop codons, whereas Chen's and Huang's methods introduce false start/stop codons every 104 and 5.78 × 105 nucleotide bases, respectively.

The remainder of this paper is organized as follows. Section 2 discusses the requirements of reversible DNA data hiding and analyzes the advantages and disadvantages of conventional methods. Section 3 explains the numeric coding of nucleotide bases, the prevention of false start/stop codons, and the NHS and CHS methods in detail. Section 4 presents and analyzes the experimental results of the proposed methods and compares them with those of conventional methods. Finally, Section 5 concludes the paper.

2. Related Works

2.1. Requirements of Reversible Data Hiding in DNA Sequence. Recently, it has been shown that the genome of a genetically tractable organism can be used as a medium for data hiding depending on the required application [1–14]. The coding DNA of a gene is transcribed by codons of three nucleotide bases that specify amino acids to encode proteins. The watermark should be embedded into coding DNA by considering codon degeneracy, to preserve the protein sequence, and codon optimization, including codon usage and GC content [10, 11]. Codon degeneracy and optimization make it difficult to embed a large number of bits into coding DNA. Noncoding DNA was initially thought to lack biological function and was referred to as "junk DNA." However, it is now clear that some noncoding DNA act as genetic switches that regulate gene expression and determine the levels or location of expression of various genes via transcription factor binding. Data hiding in noncoding DNA may damage unknown genes or gene regulatory networks. Heider et al. [9] experimentally showed that an integrated watermark deactivates the lac promoter and the RNA molecules display altered configuration after watermark introduction. Thus, they did not recommend integrating a watermark sequence into a noncoding regulatory sequence. Therefore, noncoding DNA in nonliving organisms or primitive organisms, such as bacteria, is suitable for DNA storage [1–3], DNA steganography [4–6], reversible DNA data hiding [17–20, 30–32], or fragile DNA watermarking requiring high data capacity.

There are several considerations for reversible data hiding in noncoding DNA.

(1) Dynamic Range. Nucleotide bases are described by one of four-character symbols (A, T, C, and G (DNA) or U (RNA)) with 2-bit representation. Compared to 8-bit or 10-bit image data, the 2-bit capacity for nucleotide bases is extremely low for high-capacity watermarking. A combination of nucleotide bases should be used to increase the dynamic range for more effective processing. For example, a series of four nucleotide bases can be coded with 8-bit values (44 = 256 levels).

(2) False Start/Stop Codon. The watermark can change any nucleotide base in noncoding DNA to a start codon ("ATG" (Methionine)) or three stop codons ("TAG" (Amber), "TAA" (Ochre), "TGA" (Opal)), indicating the start/stop of the coding DNA region [11]. Here, we refer to this as a "false

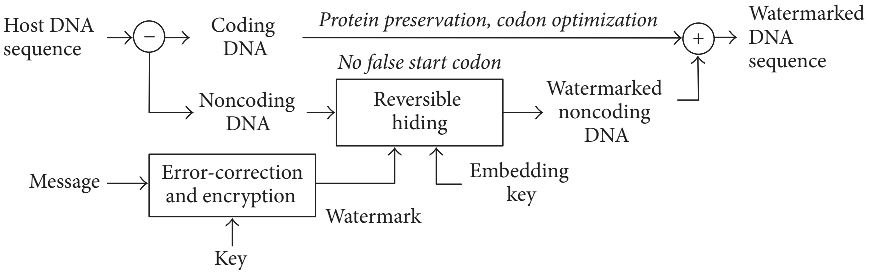


Figure 1: General process for reversible data hiding in noncoding DNA.

start/stop codon." Because these false codons have fatal effects on biological function, they should be prevented during the embedding process.

(3) Blindness without Appending. Blind detection and recovery, which detects the watermark and recovers the original DNA sequence without using the original DNA sequence or a reference sequence, should be practicable while preserving the length of the DNA sequence.

Figure 1 shows the general process of reversible data hiding in a noncoding DNA sequence. First, nucleotide bases in noncoding regions are coded into numerical values. The noncoding regions can be selected for embedding of the target depending on their length. The watermark coded by error correction or encryption is embedded into any or all numerical values using any reversible method while considering blind detection/recovery, biological function preservation, and capacity. The watermarked numerical values are reverse-coded into nucleotide bases in order to prepare watermarked noncoding regions. Here, two secret keys can be used for the watermark message or for embedding.

2.2. Conventional Methods. DNA steganography [4–6] and DNA watermarking [7–14] for DNA data hiding have been investigated in numerous studies. However, not all of these methods are reversible. We now consider some reversible data hiding methods.

Chen [19] adopted lossless compression and difference expansion (DE), which has been widely used for reversible image watermarking [21]. They coded a sequence of 2-bit binary "ATCG" values into decimal values as a unit of $|w|$ bits, classified pairs of decimal values into expandable set S1 and changeable set S2, appended the compressed location map, original LSBs of pairs in S2 (LSB(S2)), and secret binary message in the last compressed values, and embedded them into differences of pairs in S1 using difference expansion as described by Tian [21]. They experimentally demonstrated a payload capacity of 0.09–0.13 bpn with $|w| = 2$, which is extremely low. This method does not embed sufficient payload and does not consider preventing false start/stop codons.

Huang et al. [20] presented histogram-based reversible data hiding for low modification of nucleotide bases. This method generates a histogram of decimal values that are coded by every $2t$ bits of consecutive nucleotide bases. Note that h is the most frequent value, $L1$ is the least frequent value, and $L2$ is the second-least frequent value in the histogram; p is a decimal value; and b is a watermark bit. If $p = L1$, then set p to $L2$ and the location map to 1. If $p = L2$, then set the location map to 0 without changing p. If $p = h$ and $b = 0$, then set p to $L1$. Otherwise, if $p = h$ and $b = 1$, p remains unchanged. Detection and recovery are performed using h, $L1$, $L2$, and the location map. Huang et al. experimentally showed that the watermark capacity and the modification rate were 0.024 bpn and 4.07%–4.80% with $t = 2$ and 0.011 bpn and 1.86%–2.34% with $t = 3$. Although this method achieves a low modification rate, the watermark capacity is extremely low and false start/stop codons are produced, as in the case of Chen's method.

Lee and Kwon applied CDE-MBE [17] and LS-PE [18] to DNA code values, which allow for maximum permissible expansion within the range where no false start/stop codons are generated. CDE-MBE embeds multiple watermark bits in the maximum allowable difference expansion of the previous embedded code value and the current code value, while LS-PE embeds multiple watermark bits with the maximum allowable prediction error expansion of two code values. Both methods substitute extra information for detection and recovery with the LSB of the watermarked DNA code values. It was reported that the watermark capacities of LS-PE and CDE-MBE are 0.419 bpn and 0.235 bpn, respectively, on average [18]. Thus, although LS-PE has a higher watermark capacity, it has low capacity efficiency. Its ratio of watermark data to extra data is 23.4%; hence, the extra data required by this method is 4.3 times greater than the watermark data.

Liu et al. [31] presented the piecewise linear chaotic map- (PWLCM-) based reversible data hiding method for DNA sequences. Fu et al. [32] and Ma et al. [33] presented reversible data hiding methods for tamper location and tamper restoration of DNA sequences via substitution by the complementary rule. However, these methods are nonblind in that the detection and recovery processes require the original DNA sequence or a reference sequence.

Most conventional methods do not prevent the introduction of false start/stop codon and have low capacity efficiency. In this paper, we describe reversible data hiding methods that produce no false start/stop codons while achieving not only blind detection and recovery but also high watermark capacity.

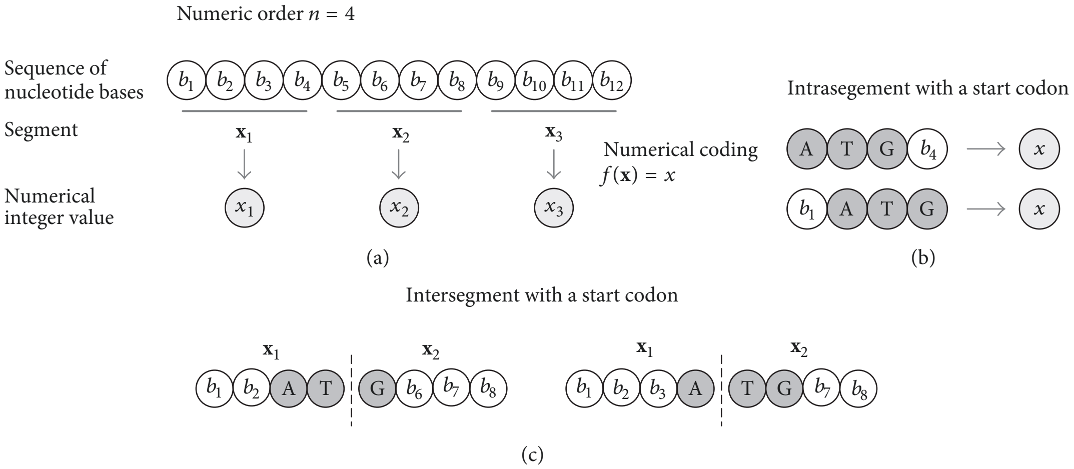

Numeric order n = 4
Sequence of nucleotide bases
b1 b2 b3 b4 b5 b6 b7 b8 b9 b10 b11 b12
Segment
x1
x2
x3
Numerical coding f(x) = x
Numerical integer value
x1
x2
x3
(a)
Intrasegement with a start codon
A T G b4 → x
b1 A T G → x
(b)
Intersegment with a start codon
x1 x2
b1 b2 A T G b6 b7 b8
x1 x2
b1 b2 b3 A T G b7 b8
(c)

FIGURE 2: (a) 8-bit value x for a segment $\mathbf{x}$ of 4 nucleotide bases (numeric order $n = 4$) and false start codon occurrence (b) in a segment (intrasegment) and (c) between segments (intersegment).

3. Proposed Reversible Data Hiding on DNA Sequence

This section introduces reversible data hiding methods using multilevel histogram shifting with the following features. (1) *Blindness*: the watermark is embedded without changing the sequence length or amino acids. Further, the watermark is detected and the original DNA sequence is recovered without the need for a reference DNA sequence. (2) *Numerical facility*: the watermarking process is facilitated by encoding consecutive nucleotide bases into $2n$-bit integer values. (3) *High capacity*: multiple bits are embedded in each integer value using multilevel histogram shifting. (4) *No false start/stop codon*: false start/stop codons are prevented via false codon searching both in an integer value and between consecutive integer values.

3.1. Preprocessing: Numerical Coding and False Codon Prevention.

Nucleotide bases of four-character symbols should be encoded into integer values to facilitate watermarking.

Numerical Coding of Nucleotide Bases. In general, a nucleotide base b of four-character symbols is represented by a 2-bit value b as shown in Figure 2(a). For example,

$$b = (\mathrm{A}, \mathrm{T}, \mathrm{C}, \mathrm{G}) \longleftrightarrow b = (0, 1, 2, 3)_{10}. \quad (1)$$

To extend the numerical value range of nucleotide bases, we encode a segment $\mathbf{x} = \{b_1, \ldots, b_n\}$ with n nucleotide bases into a $2n$-bit value x:

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

Furthermore, we can obtain nucleotide bases in $\mathbf{x}$ from x as follows:

$$b_k = (x \gg 2(n-k)) \,\%4 \quad \text{for } k = 1, \ldots, n. \quad (3)$$

$\gg$ denotes the right bit-shifting operator, and n is the numeric order, which is the number of nucleotide bases for a segment.

False Start/Stop Codon Prevention. Any false start/stop codon can be generated in a segment (intrasegment) or between segments (intersegment) by the watermark. Hereafter, for brevity, we refer to a false start/stop codon as a false codon. We prevent false codons in noncoding regions using false codon searching.

(1) Intrasegment. Given a segment $\mathbf{x}$ with n nucleotide bases, any false codon can occur $(n-2)$ times in a segment, as shown in Figure 2(b). A false codon that occurs in any position is encoded into $2^{2(n-3)}$ values. Thus, one false start codon and three false stop codons occur as a total of $4(n-2) \times 2^{2(n-3)}$ values. We generate the false codon table of all values including one false start codon or three false stop codons; then, we determine whether the watermarked value x' is included in the false codon table in the embedding process. Table 1 shows an example of the false codon table when the numeric order n is 4.

(2) Intersegment. Given adjacent watermarked segments $(\mathbf{x}'_{j-1}, \mathbf{x}'_j)$, any false codon may be present between $\mathbf{x}'_{j-1}$ and $\mathbf{x}'_j$. For example, the last nucleotide base or last two nucleotide bases in $\mathbf{x}'_{j-1}$ and first two nucleotide bases or first nucleotide base in $\mathbf{x}'_j$ can be $(\cdots\mathrm{A}, \mathrm{TG}\cdots)$ or $(\cdots\mathrm{AT}, \mathrm{G}\cdots)$ in the false start codon ATG, as shown in Figure 2(c). Adjacent watermarked segments $(\mathbf{x}'_{j-1}, \mathbf{x}'_j)$ including a false codon can be determined by searching for a value that concatenates the $n-1$th and nth nucleotide bases of $\mathbf{x}'_{j-1}$ and the first and second nucleotide bases of $\mathbf{x}'_j$ in the false codon table.

In the embedding process, we decrease the number of embeddable bits until the false codons do not occur in a segment and between adjacent segments, as explained in Sections 3.2 and 3.3.

Table 1: False codon table of integer values including false start/stop codons when the numeric code is 4 and (A, T, C, G) = (0, 1, 2, 3).

Category	Segment including false codon	Numeric coding of segment	False codon table of integer value
Start codon	ATGX	$0 \times 2^6 + 1 \times 2^4 + 3 \times 2^2 + (0, 1, 2, 3) \times 1$	28, 29, 30, 31
	XATG	$(0, 1, 2, 3) \times 2^6 + 0 \times 2^4 + 1 \times 2^2 + 3 \times 1$	7, 71, 135, 199
Stop codon	TAGX	$1 \times 2^6 + 0 \times 2^4 + 3 \times 2^2 + (0, 1, 2, 3) \times 1$	76, 77, 78, 79
	XTAG	$(0, 1, 2, 3) \times 2^6 + 1 \times 2^4 + 0 \times 2^2 + 3 \times 1$	19, 83, 147, 211
	TAAX	$1 \times 2^6 + 0 \times 2^4 + 0 \times 2^2 + (0, 1, 2, 3) \times 1$	64, 65, 66, 67
	XTAA	$(0, 1, 2, 3) \times 2^6 + 1 \times 2^4 + 0 \times 2^2 + 0 \times 1$	16, 80, 144, 208
	TGAX	$1 \times 2^6 + 3 \times 2^4 + 0 \times 2^2 + (0, 1, 2, 3) \times 1$	112, 113, 114, 115
	XTGA	$(0, 1, 2, 3) \times 2^6 + 1 \times 2^4 + 3 \times 2^2 + 0 \times 1$	24, 88, 152, 216

3.2. Noncircular Histogram Shifting (NHS) Based Reversible Data Hiding. $2n$-bit values for all segments are shifted to other values except for values in the false codon table. Let us consider multilevel histogram shifting of noncircular type.

Embedding Process. Let N be the maximum number of shifting levels for a value. We divide the range of $2n$-bit values into a number of regions having $(2N - 1)$ values such that a region P_i has bilateral symmetry with subregion P_i^{L} of left N values and subregion P_i^{R} of right N values about a center value r_i. The center value r_i is used as the reference value for multilevel shifting. Here, a residual region of values that are not included in the regions exists. This region is not selected for embedding.

Given a value x for a segment $\mathbf{x}$, a previously watermarked segment $\mathbf{x}'_{-1}$, and a watermark W, the maximum number of embeddable bits in a value is $\lceil \log_2 N \rceil$. First, we find the center value r_i of a region to which x belongs, and we determine the number of embeddable bits k on the basis of the difference $d = x - r_i$:

$$k = \begin{cases} \lceil \log_2 N \rceil - (j - 1), & \text{if } 2^{j-1} \le |d| < 2^j \\ 0, & \text{if } |d| = 0. \end{cases} \tag{4}$$

If x is a center value r_i, k is 0, that is, no bits are selected for embedding. Otherwise, we shift x by up to k bits of watermark $\{w_j : w_j \in W,\ 1 \le j \le k\}$ while checking whether x' does not include false codons:

$$x' = r_i + 2^k d + \operatorname{sgn}(d) \sum_{j=1}^{k} 2^{j-1} w_j. \tag{5}$$

If a shifted value x' is in the false codon table, we decrease k by 1 and shift x by up to $k - 1$ bits. We obtain a sequence of all watermarked segments $\mathbf{X}'$ by repeating this process until $k = 0$ for all segments.

Our approach requires an extra dataset $S = K \cup T \cup B$ for detection and recovery, including a set K of the numbers of embedded bits in segments, a set T of the region markers of shifted center values, and a set B of binary LSBs of nucleotide bases in $\mathbf{X}'$. The extra dataset is compressed losslessly and the compressed extra dataset S' is substituted into binary LSBs of nucleotide bases in $\mathbf{X}'$. A DNA sequence $\mathbf{D}'$ with watermarked segments $\mathbf{X}''$ including S' is transmitted or stored.

Histogram Shifting. The region P_i of a histogram is divided into a left subregion P_i^{L} and a right subregion P_i^{R} by a center value r_i. Figure 3 shows the multilevel shifting of values $\{x : x \in P_i\}$ by the difference $|d|$ with r_i and watermark bits when the maximum number of embeddable bits is 3. The values with $|d|$ equal to 1 can be shifted by up to 3 bits ($k = 3$), while those with $|d|$ between 4 and 7 can be shifted by up to 2 bits ($k = 2$) and those with $|d|$ between 2 and 3 can be shifted by up to 1 bit ($k = 1$). The value $x = r_i$ of ($|d| = 0$) is fixed.

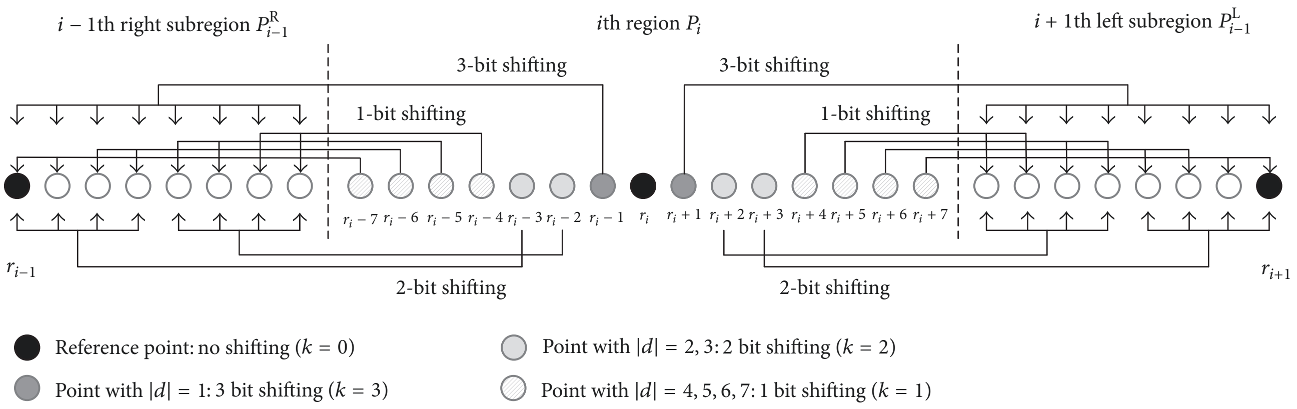
The values in the right subregion P_i^{R} are shifted to the next left subregion P_{i+1}^{L}. By contrast, the values in the left subregion P_i^{L} are shifted to the previous right subregion P_{i-1}^{R}, as shown in Figure 4(a).

The shifted center value can be observed in three cases. The first case is that a value x is the same as r_i, which is fixed. The other two cases include values in P_{i-1}^{R} or P_{i+1}^{L} shifted to r_i. The number of embedded bits k indicates whether the original value is r_i or whether it is shifted to r_i from the left or right regions. Therefore, the region mark τ of the shifted center values indicates where it is shifted from P_{i-1}^{R} or P_{i+1}^{L} for the detection and recovery process.

$$\tau = \begin{cases} 0, & \text{if } x' = r_i,\ x \in P_{i-1}^{\mathrm{R}} \\ 1, & \text{if } x' = r_i,\ x \in P_{i+1.}^{\mathrm{L}} \end{cases} \tag{6}$$

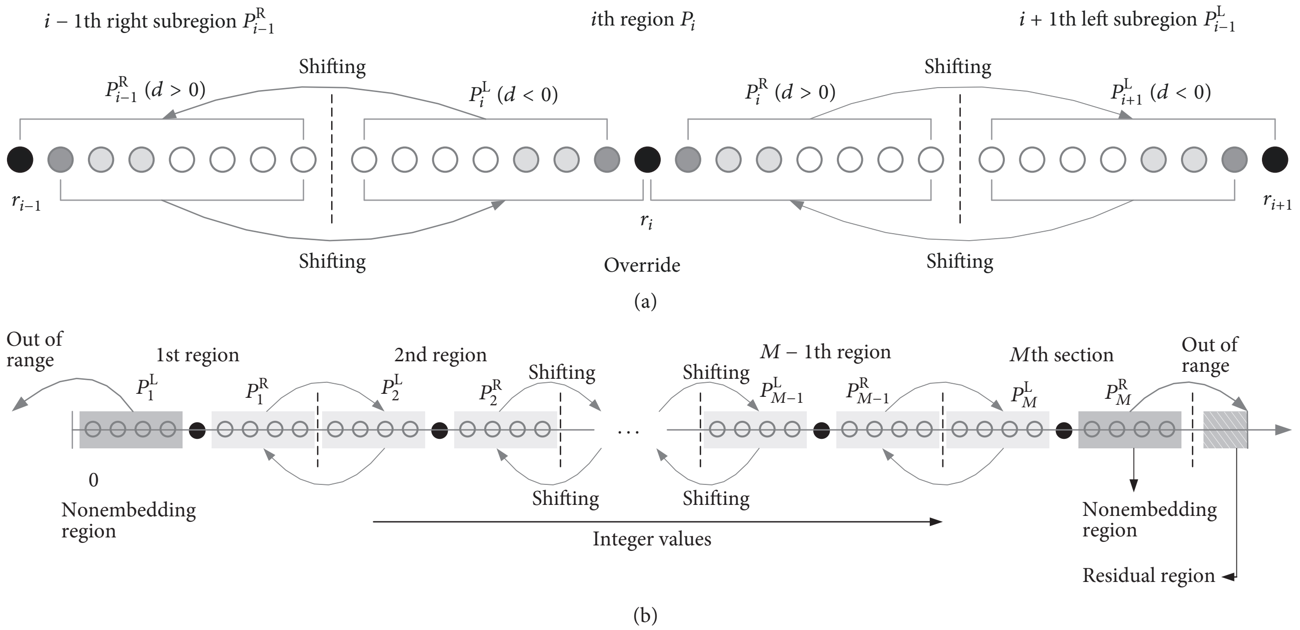
Figure 4(b) shows that subregions from P_1^{R} to P_M^{L} among a total of M regions are shifted toward each other except for the two boundary subregions P_1^{L} and P_M^{R}, which are nonembedding regions.

Detection and Recovery Process. We extract the compressed extra dataset S' from binary LSBs B' of nucleotide bases in transmitted segments $\mathbf{X}''$ including S' and decompress S' to obtain the extra dataset $S = K \cup T \cup B$. Here, the watermarked segments $\mathbf{X}'$ can be easily obtained by substituting B and B'. Next, we detect the watermark W and recover the original values X of segments $\mathbf{X}$ by a set K of the numbers of embedded bits and a set T of the region markers of shifted center values.



$i-1$th right subregion P_{i-1}^{R}
ith region P_i
$i+1$th left subregion P_{i-1}^{L}
3-bit shifting
3-bit shifting
1-bit shifting
1-bit shifting
r_i-7 r_i-6 r_i-5 r_i-4 r_i-3 r_i-2 r_i-1 r_i r_i+1 r_i+2 r_i+3 r_i+4 r_i+5 r_i+6 r_i+7
r_{i-1}
r_{i+1}
2-bit shifting
2-bit shifting
● Reference point: no shifting ($k=0$)
● Point with $|d|=1$: 3 bit shifting ($k=3$)
○ Point with $|d|=2,3$: 2 bit shifting ($k=2$)
▨ Point with $|d|=4,5,6,7$: 1 bit shifting ($k=1$)

FIGURE 3: Multilevel shifting of values with difference d and center value r_i where $|d|>0$ when the maximum number of embeddable bits is 3 for a histogram region P_i.



(a)

(b)

FIGURE 4: (a) Shifting of two subregions P_i^{R} and P_{i+1}^{L} and that of P_i^{L} and P_{i-1}^{R} and (b) shifting of two adjacent subregions of different regions except for two boundary subregions.

Given a value x' of a watermarked segment $\mathbf{x}'$ and the number of embedded bits k of $\mathbf{x}'$, we first obtain the center value r of the region to which the original value x belongs. This can be determined by the region P_i of x' and the region marker τ in both cases:

$$r=\begin{cases} r_{i-1}, & \text{if } x' \in P_i^{\mathrm{L}} \text{ or } \left(x'=r_i,\ \tau=0\right) \\ r_{i+1}, & \text{if } x' \in P_i^{\mathrm{R}} \text{ or } \left(x'=r_i,\ \tau=1\right). \end{cases} \tag{7}$$

When x' is not the center value r_i of P_i, the region of x can be easily determined as the left subregion P_{i-1} or right subregion P_{i+1} according to $x' \in P_i^{\mathrm{L}}$ or $x' \in P_i^{\mathrm{R}}$, respectively. However, when $x' \in P_i$ and $x'=r_i$, the region of x can be known as P_{i-1} or P_{i+1} by the region marker τ. Based on the center value r, k-bit watermark $\{w_j : w_j \in W,\ 1 \le j \le k\}$ is detected and the original value x is recovered as follows:

$$\begin{aligned} w_j &= \left(\left(x'-r\right) \gg (j-1)\right) \% 2 \quad \text{for } j=1,\ldots,k, \\ x &= r+\left(\left(x'-r\right) \gg k\right). \end{aligned} \tag{8}$$

Watermark Capacity and Extra Data Capacity. Balado [15, 16] modeled the Shannon watermark capacity for noncoding and coding regions under mutations. Considering the capacity of a noncoding region, 2 bpn can always be embedded into a 4-ary alphabet of a nucleotide base without any mutations. In the case of substitution mutation, the capacity of a 4-ary symmetric channel, which is achieved for a uniform input, is $2+(1-q)\log(1-q)q\log(q/3)$ bpn, where q is the substitution probability of the nucleotide base. For example, if 25% of nucleotide bases are substituted, the maximum capacity is approximately 1.636 bpn. Thus, Balado analyzed the capacity of irreversible watermarking without considering reversible watermarking. Reversible watermarking, which can detect the watermark while recovering the original sequence from the watermarked sequence, has two additional constraints of reversibility and extra data. Therefore, we analyze the capacity of our method under reversibility and extra data rather than following Balado's capacity analysis.

We compute the watermark bpn and the extra dataset bpn to analyze the capacity performance. The watermark bpn is defined as the watermark bits embedded in one nucleotide base, denoted by bpn^{W}, and the extra dataset bpn is defined as the additional data bits for a nucleotide base, denoted by $\mathrm{bpn}^{\mathrm{Extra}}$. The extra dataset is required to detect the watermark and recover the original sequence. The watermark bpn should be high whereas the extra dataset bpn should be low.

Let $C(P_i^{\mathrm{L}})$ and $C(P_i^{\mathrm{R}})$ be the number of embedded bits in two subregions P_i^{L} and P_i^{R}. The total number of embedded bits is the sum of embedded bits in all subregions except the two boundary subregions P_1^{L} and P_M^{R}. Thus, the number of watermark bits per nucleotide base $\mathrm{bpn}_{\mathrm{NHS}}^{W}$ is given by

$$\mathrm{bpn}_{\mathrm{NHS}}^{W} = \frac{1}{n\,|\mathbf{X}|}\left(C\left(P_1^{\mathrm{R}}\right) + \sum_{i=2}^{M-1}\left(C\left(P_i^{\mathrm{L}}\right) + C\left(P_i^{\mathrm{R}}\right)\right) + C\left(P_M^{\mathrm{L}}\right)\right)[\text{bits/base}], \tag{9}$$

where $|\mathbf{X}|$ is the total number of segments and n is the number of nucleotide bases in a segment. Thus, $n|\mathbf{X}|$ denotes the total number of nucleotide bases in all segments.

Let $\mathrm{Extra}_{\mathrm{NHS}}$ be the number of bits for storing uncompressed extra dataset $S = K \cup T \cup B$ required for detection and recovery. Since the maximum number of embeddable bits in a value is $\lceil \log_2 N \rceil$, it is represented by $\lceil \log_2(\lceil \log_2 N \rceil) \rceil$ bits. Therefore, a set K of the numbers of embedded bits for all values is represented by $\lceil \log_2(\lceil \log_2 N \rceil) \rceil \times |\mathbf{X}|$ bits. A region mark for the shifted center value can be stored by one bit that indicates whether it is shifted from the left region or the right region plus the position information on a sequence of segments, which can be represented by $\lceil \log_2 |\mathbf{X}| \rceil$ bits. Let $|X' = R|$ be the total number of shifted center values. A set T of region marks for shifted center values is represented by $\lceil \log_2 |\mathbf{X}| \rceil \times |X' = R|$ bits. The size of a set B of LSBs of binary nucleotide bases in $\mathbf{X}'$ is equal to the total number of nucleotide bases. Thus, B is represented by $|\mathbf{X}|$ bits. In summary, the uncompressed extra dataset $\mathrm{Extra}_{\mathrm{NHS}}$ is represented as follows:

$$\mathrm{Extra}_{\mathrm{NHS}} = \lceil \log_2(\lceil \log_2 N \rceil) \rceil \times |\mathbf{X}| + \lceil \log_2 |\mathbf{X}| \rceil \times \left|X' = R\right| + |\mathbf{X}|\,[\text{bits}]. \tag{10}$$

Let ρ be the lossless compression ratio. The number of bits of the compressed extra dataset per nucleotide base, $\mathrm{bpn}_{\mathrm{NHS}}^{\mathrm{Extra}}$, can be computed as follows:

$$\mathrm{bpn}_{\mathrm{NHS}}^{\mathrm{Extra}} = \frac{\rho}{|\mathbf{X}|}\mathrm{Extra}_{\mathrm{NHS}}\,[\text{bits/base}]. \tag{11}$$

3.3. Circular Histogram Shifting (CHS) Based Reversible Data Hiding. Unlike image quality, integer values of DNA segments can be shifted to any values. In addition, maximum and minimum values can be shifted toward each other only if the condition of false codon is satisfied. The CHS-based method makes the histogram domain circular so that the two boundary subregions P_1^{L} and P_M^{R} can be shifted toward each other.

Embedding Process. As in the previous subsection, we divide the range of $2n$-bit values into M regions $\{P_i : 1 \le i \le M\}$, and each range consists of $(2N - 1)$ values. Here, the residual region of $[M(2N - 1) + 1, 2^{2n} - 1]$ is generated, as shown in Figure 5. Because the residual region exists between P_1^{L} and P_M^{R}, it is difficult to shift the values in the two subregions toward each other. Therefore, we exchange the right subregion P_M^{R} of the last region and the residual region. Here, the detached region P_M has two center values of subregions individually: r_M^{R} for P_M^{R} and r_M^{L} for P_M^{L}.

Given a value x, we find the center value r_i of the region to which x belongs; then, we shift x by up to k bits of watermark $\{w_j : w_j \in W, 1 \le j \le k\}$, which is determined by the difference $d = k - r_i$.

$$x' = \left(r_i + 2^k + \operatorname{sgn}(d)\sum_{j=1}^{k} 2^{j-1} w_j\right) \% 2^{2n}. \tag{12}$$

Here, values in the exchanged residual region and center values in each region are excluded from the embedding. The region marker τ of the center values shifted from an adjacent region is set as follows:

$$\tau = \begin{cases} 0, & \text{if } \left(x' = r_i,\ x \in P_{i-1}\right) \text{ or } \left(x' = r_M^{\mathrm{R}},\ x \in P_1\right) \\ 1, & \text{if } \left(x' = r_i,\ x \in P_{i+1}\right) \text{ or } \left(x' = r_1,\ x \in P_M^{\mathrm{R}}\right). \end{cases} \tag{13}$$

We obtain watermarked segments $\mathbf{X}'$ by shifting as many watermark bits as possible into all values except for the residual region while preventing false codons. The extra dataset required for the detecting and recovering process is $S = K \cup T \cup B$, which is the same as the extra dataset for the NHS-based method. We substitute the compressed extra dataset S' into LSBs of binary nucleotide bases in $\mathbf{X}'$ and then obtain a DNA sequence with watermarked segments $\mathbf{X}''$ including the compressed extra dataset.

Detection and Recovery Process. As with the NHS-based method, we obtain watermarked segments $\mathbf{X}'$ from a transmitted DNA sequence via LSB substitution of the compressed extra dataset S' and then detect the watermark W and recover

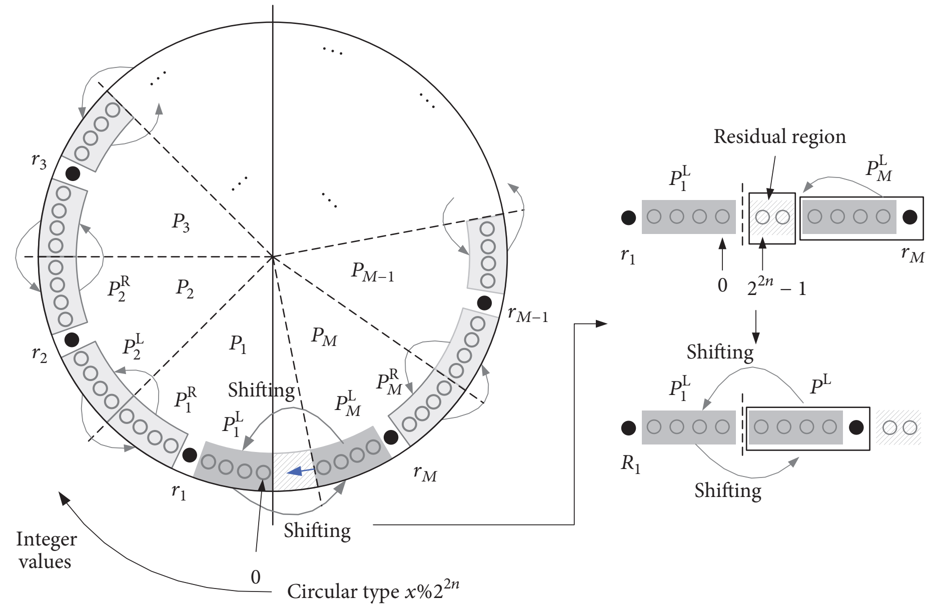


Figure 5: Circular histogram and shifting two boundary subregions toward each other in CHS-based method.

the original values from the watermarked values using the extra dataset.

Given a watermarked value x' in $\mathbf{X}'$, we find the center value r of the nonboundary region or boundary region to which x' belongs:

$$r = \begin{cases} r_{i-1}, & \text{if } x' \in P_i^{\mathrm{L}} \text{ or } \left(x' = r_i,\ \tau = 0\right) \\ r_{i+1}, & \text{if } x' \in P_i^{\mathrm{R}} \text{ or } \left(x' = r_i,\ \tau = 1\right) \end{cases}$$

$$\text{for non-boundary regions;} \tag{14}$$

$$r = \begin{cases} r_M^{\mathrm{R}}, & \text{if } 0 \le x' \le r_1 \text{ or } \left(x' = r_i,\ \tau = 0\right) \\ r_1, & \text{if } r_M^{\mathrm{R}} \le x' \le 2^{2n-1} \text{ or } \left(x' = r_M^{\mathrm{R}},\ \tau = 1\right) \end{cases}$$

$$\text{for boundary regions.}$$

Then, we detect k bits of the watermark and recover the original value x using the center value r.

$$w_j = \left(\left(\left(x' - r\right)\%2^{2n}\right) \gg (j-1)\right)\%2 \qquad \text{for } j = 1, \ldots, k; \tag{15}$$

$$x = r + \left(\left(x' - r\right)\%2^{2n} \gg k\right).$$

Watermark Capacity and Extra Data Capacity. The CHS-based method embeds watermark bits into all regions except for the residual region in the histogram domain. The total number of embedded bits is the sum of embedded bits in all subregions. Thus, the number of watermark bits per nucleotide base $\mathrm{bpn}_{\mathrm{CHS}}^{W}$ is defined as follows:

$$\mathrm{bpn}_{\mathrm{CHS}}^{W} = \frac{1}{|\mathbf{X}|}\sum_{i=1}^{M}\left(C\left(P_i^{\mathrm{R}}\right) + C\left(P_i^{\mathrm{L}}\right)\right) \text{ [bits/base]}. \tag{16}$$

Let $\mathrm{Extra}_{\mathrm{CHS}}$ be the number of bits of uncompressed extra dataset S for the CHS-based method. It can be represented in the same way as $\mathrm{Extra}_{\mathrm{NHS}}$ of the NHS-based method. Based on the lossless compression ratio ρ, the number of bits of the compressed extra dataset per nucleotide base, $\mathrm{bpn}_{\mathrm{CHS}}^{\mathrm{Extra}}$, can be computed as follows:

$$\mathrm{bpn}_{\mathrm{CHS}}^{\mathrm{Extra}} = \frac{\rho}{|\mathbf{X}|}\mathrm{Extra}_{\mathrm{CHS}} \text{ [bits/base]}. \tag{17}$$

The extra datasets for the CHS-based and NHS-based methods are the same, but the watermark capacity of the CHS-based method is higher than that of the NHS-based method.

4. Experimental Results

Image quality in terms of PSNR versus bits per pixel is the main evaluation metric for reversible watermarking. However, DNA reversible data hiding methods should be evaluated on the basis of the preservation of biological function versus bpn. Our methods do not introduce any false start/stop codons in noncoding regions while preserving the amino acid code. We evaluated the watermark bpn bpn^{W}, the extra data bpn $\mathrm{bpn}^{\mathrm{Extra}}$, the base change rate e, and the occurrence of false codons using our NHS- and CHS-based methods as well as Chen's method [19], Huang's method [20], and LS-PE [18]. Here, the base change rate E is the rate of change of nucleotide bases by watermark bits. Given the original segments $\mathbf{X}$ and watermarked segments $\mathbf{X}'$, the base change rate E is defined as follows:

$$E = \frac{1}{n|\mathbf{X}|}\sum_{i=1}^{|\mathbf{X}|}\sum_{j=1}^{n} e_{ij} \quad \text{where } e_{ij} = \begin{cases} 1, & \text{if } b_{ij} \ne b'_{ij} \\ 0, & \text{if } b_{ij} = b'_{ij}. \end{cases} \tag{18}$$

Table 2: Test DNA sequences which are reused from [17].

Type	Access number	Total bases	Number of noncoding regions	Number of noncoding DNA bases
Archaea	AE017199	490,885	289	38,932
Bacterium	CP000108	2,572,079	1,770	301,761
Bacterium	CP000247	4,938,920	3,850	570,214
Bacterium	CP000672.1	1,887,192	1,444	466,266
Bacterium	AF012886.2	6,756	7	2,058
Bacterium	AE014075.1	5,231,428	3,767	631,026
Bacterium	CP000473.1	9,965,640	6,224	962,527
Eukaryota	nm_000520	2,437	2	847
Eukaryota	NC_001709.1	19,517	11	8,347
Eukaryota	NC_006033	1,195,132	533	393,739
Eukaryota	AL161582.2	198,669	302	137,622
Eukaryota	AL161595.2	198,151	215	126,917
Eukaryote	NC_006047	2,007,515	1,099	516,557
Moss	AP005672.1	122,890	99	51,916
Plant	NC_025652.1	141,255	68	91,971
Virus	AY653733.1	1,181,404	883	155,805

Assuming that nucleotide bases changed by a random watermark are uniformly distributed, the base change rate was nearly 3/4 = 0.75.

The watermark bpn and extra data bpn of LS-PE, $\text{bpn}_{\text{PE}}^{W}$ and $\text{bpn}_{\text{PE}}^{\text{Extra}}$, respectively, depend on the numeric order n and the prediction order p. By contrast, the watermark bpn and extra data bpn of the NHS- and CHS-based methods, $\text{bpn}_{\text{NHS}}^{W}$, $\text{bpn}_{\text{CHS}}^{W}$ and $\text{bpn}_{\text{NHS}}^{\text{Extra}}$, $\text{bpn}_{\text{CHS}}^{\text{Extra}}$, respectively, depend on the numeric order n and the maximum embeddable bit number $k_{\max}$. Therefore, we experimentally selected parameters with the most watermark bpns in each method and used them to compare the capacities of our methods with those of the conventional methods. Specifically, we used $|w| = 2$ in Chen's method and $|t| = 2$ in Huang's method.

We used test DNA sequences provided by NCBI GenBank that are the same experimental sequences of [17]. Table 2 summarizes the type, access number, total number of nucleotide bases, number of noncoding regions, and number of nucleotide bases in the noncoding regions of our test DNA sequences. The test DNA sequences varied in length. Noncoding regions with a small number of nucleotide bases were not used for the embedding regions.

4.1. Parameter Setting. Here, we discuss how to determine the parameters of our methods for comparison with the conventional methods. Given the numeric order n, the maximum embeddable bit number $k_{\max}$ for each region is $k_{\max} \leq 2n-2$. Figure 6 shows the extra data bpn versus the watermark bpn and the base change rate versus the watermark bpn for the NHS- and CHS-based methods, where $k_{\max}$ is varied from 2 to $2n-2$ in $n \in [2, 10]$. The watermark bpn is the highest when $n = 2$ and $k_{\max} = 2$. In this case, $\text{bpn}_{\text{CHS}}^{W}$ is 0.175 bpn higher than $\text{bpn}_{\text{NHS}}^{W}$ ($\text{bpn}_{\text{CHS}}^{W} = 0.566$ bpn and $\text{bpn}_{\text{NHS}}^{W} = 0.391$ bpn). For a given n, $\text{bpn}_{\text{CHS}}^{W}$ increases with $k_{\max}$. As n increases, $\text{bpn}_{\text{CHS}}^{W}$ and $\text{bpn}_{\text{NHS}}^{W}$ decrease. Based on these results, $\text{bpn}_{\text{CHS}}^{W}$ is the highest when $k_{\max} = 2n - 2$ and $\text{bpn}_{\text{NHS}}^{W}$ is the highest when $k_{\max} = \lfloor 1/4(5n - 2) \rfloor$.

With regard to the extra data bpn versus the watermark bpn, the two bpns decrease if the numeric order n increases. For $n = 2$ and $k_{\max} = 2$, the CHS-based method requires extra data bpn $\text{bpn}_{\text{CHS}}^{\text{Extra}} = 0.303$ bpn for $\text{bpn}_{\text{CHS}}^{W} = 0.566$ bpn and the NHS-based method requires extra data bpn $\text{bpn}_{\text{NHS}}^{\text{Extra}} = 0.315$ bpn for $\text{bpn}_{\text{NHS}}^{W} = 0.391$ bpn. These extra data are approximately one-third of the LSB substitutable bits, for which we can substitute the extra data into the LSBs of binary nucleotide bases three times. With regard to the base change rate E versus the watermark bpn, the former increases with the numeric order n. The CHS-based method has $E = 0.701$ for $n = 2$, $k_{\max} = 2$ with the highest $\text{bpn}_{\text{NHS}}^{W}$, but it has $E = 0.778$ and 0.765 for $n = 3$, $k_{\max} = 4$ and $n = 4$, $k_{\max} = 5$, respectively, with low $\text{bpn}_{\text{NHS}}^{W}$. Similarly, the NHS-based method has $E = 0.465$ for $n = 2$, $k_{\max} = 2$ with the highest $\text{bpn}_{\text{NHS}}^{W}$, but it has $E = 0.498$ and 0.474 for $n = 3$, $k_{\max} = 3$ and $n = 4$, $k_{\max} = 3$, respectively, with low $\text{bpn}_{\text{NHS}}^{W}$.

According to the two parameters n and $k_{\max}$, the watermark bpn of the CHS-based method is 0.004–0.175 higher than that of the NHS-based method under a similar quantity of extra data, while the base change rate of the former is 0.038–2.760 higher than that of the latter. On the basis of these results, we set the maximum embeddable bit number $k_{\max}$ to 20 in the NHS- and CHS-based methods and then compared them with the LS-PE-based method [18] and other methods [19, 20] by varying the numeric order n.

4.2. Comparison of Watermark Capacity, Extra Data Capacity, and Base Change Rate. We set the parameters of the LS-PE-based method ($p = 20, 30$), NHS-based method ($k_{\max} = \lfloor 1/4(5n - 2) \rfloor$), and CHS-based method ($k_{\max} = 2n - 2$) for numeric order $n \in [2, 6]$, as well as for Chen's method ($|w| = 2$) and Huang's method ($|t| = 2$), such that the

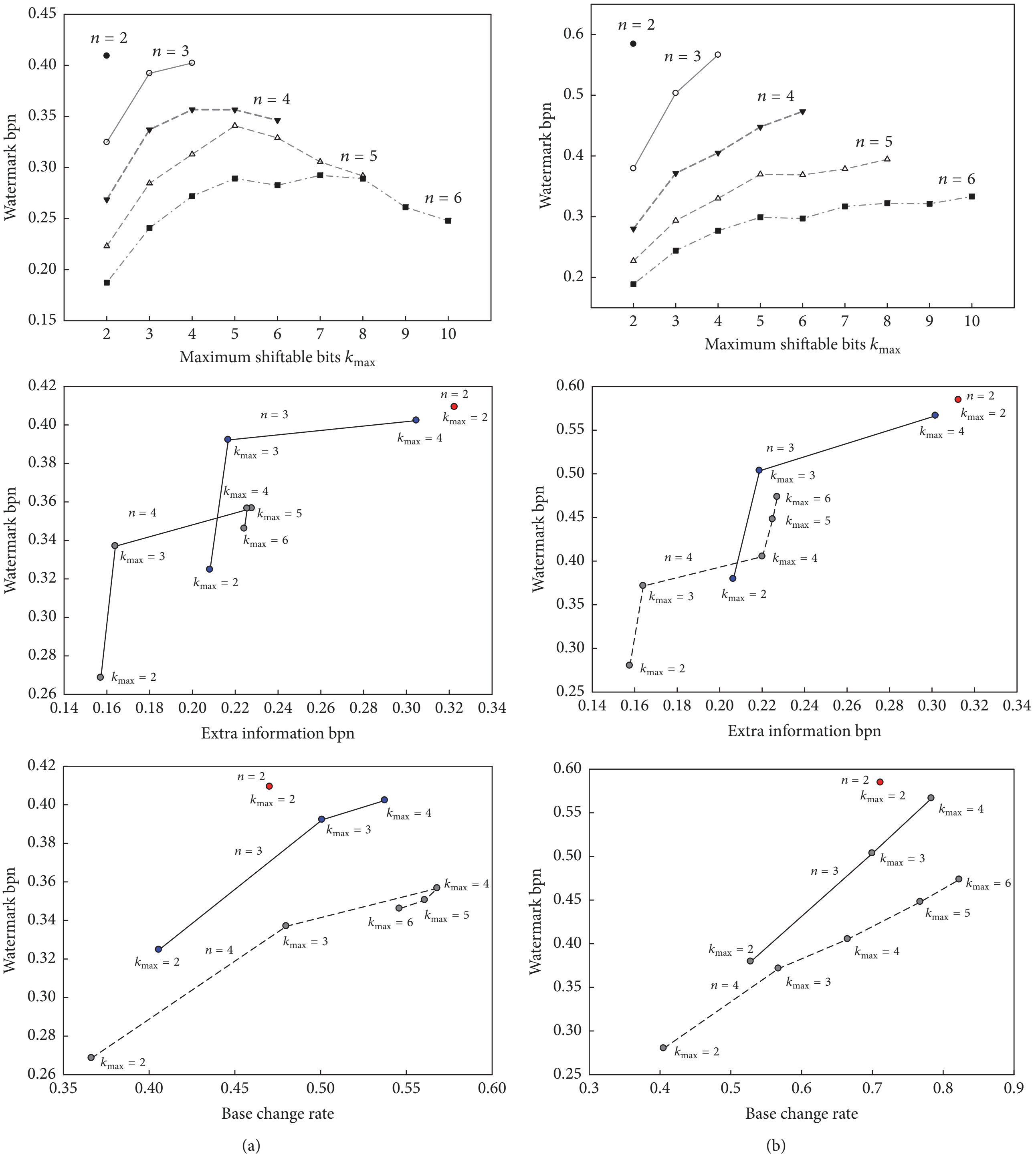
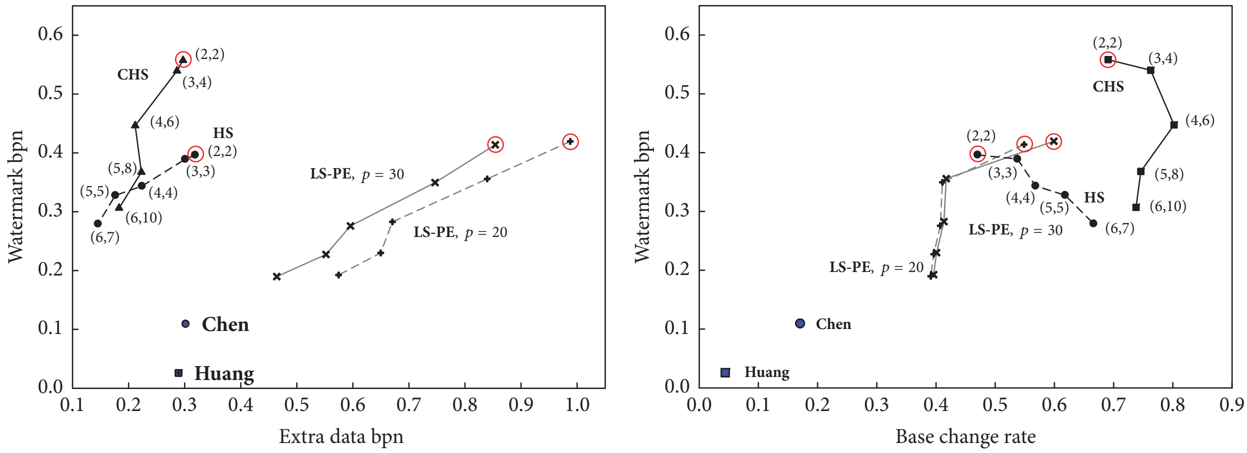


Figure 6: Watermark bpn, extra data bpn versus watermark bpn, and base change rate versus watermark bpn by varying the numeric order n and the maximum embeddable bit number k_{max} of (a) NHS-based method and (b) CHS-based method.

highest watermark bpn is achieved. Next, we compared the watermark bpn, extra data bpn, and base change rate of these methods. Figure 7 shows the extra data bpn and the base change rate versus the watermark bpn of each method.

With regard to the watermark bpn, the CHS-based method with $(n, k_{max}) = (2, 2)$ showed the highest value at 0.566 bpn. The next highest values were 0.419 bpn for the LS-PE-based method with $p = 30$, 0.413 bpn for the LS-PE-based method [18] with $p = 20$, and 0.391 bpn for the NHS-based method with $(n, k_{max}) = (2, 2)$. For Chen's [19] and Huang's [20] methods, the values were 0.108 bpn and 0.027 bpn, respectively, which are extremely low compared



(a)

(b)

Figure 7: (a) Extra data bpn versus watermark bpn and (b) base change rate versus watermark bpn of NHS- and CHS-based methods $(n, k_{\max})$ and conventional methods.

Table 3: Watermark bpn and extra data bpn for test DNA sequences (watermark bpn : embedded watermark bit per nucleotide base, extra data bpn : extra data bit per nucleotide base required for watermark detection/recovery).

Type	Access number	Proposed NHS $(n, k_{\max}) = (2, 2)$		Proposed CHS $(n, k_{\max}) = (2, 2)$		LS-PE [18] $(n, p) = (2, 30)$		Chen [19] $(\lvert w\rvert = 2)$		Huang et al. [20] $(t = 2)$	
		Watermark bpn	Extra data bpn	Watermark bpn	Extra data bpn	Watermark bpn	Extra data bpn	Watermark bpn	Extra data bpn	Watermark bpn	Extra data bpn
Archaea	AE017199	0.396	0.308	0.567	0.295	0.425	0.930	0.098	0.232	0.034	0.202
Bacterium	CP000108	0.390	0.312	0.554	0.296	0.414	1.159	0.114	0.207	0.015	0.161
Bacterium	CP000247	0.396	0.308	0.565	0.292	0.412	1.194	0.096	0.218	0.021	0.208
Bacterium	CP000672.1	0.387	0.301	0.562	0.293	0.401	0.670	0.098	0.222	0.044	0.170
Bacterium	AF012886.2	0.356	0.332	0.590	0.332	0.429	5.300	0.111	0.213	0.022	0.242
Bacterium	AE014075.1	0.404	0.305	0.567	0.291	0.404	1.023	0.115	0.216	0.044	0.207
Bacterium	CP000473.1	0.425	0.311	0.555	0.299	0.405	1.048	0.115	0.196	0.037	0.215
Eukaryota	nm_000520	0.399	0.387	0.507	0.369	0.409	10.601	0.099	0.217	0.028	0.175
Eukaryota	NC_006033	0.389	0.301	0.578	0.292	0.414	0.672	0.117	0.207	0.025	0.213
Eukaryota	AL161582.2	0.381	0.312	0.569	0.301	0.422	0.611	0.115	0.210	0.046	0.223
Eukaryota	AL161595.2	0.378	0.313	0.574	0.304	0.419	0.559	0.097	0.235	0.003	0.198
Eukaryote	NC_006047	0.402	0.308	0.581	0.291	0.411	0.793	0.115	0.217	0.016	0.221
Moss	AP005672.1	0.389	0.298	0.575	0.303	0.455	0.662	0.109	0.227	0.012	0.204
Plant	NC_025652.1	0.401	0.315	0.569	0.300	0.398	0.470	0.116	0.213	0.037	0.183
Virus	AY653733.1	0.377	0.310	0.579	0.297	0.463	1.131	0.101	0.237	0.015	0.227
Average	-	0.391	0.315	**0.566**	0.303	0.419	1.788	0.108	0.218	0.027	0.203
Capacity efficiency	Watermark/ extra data	1.243		**1.865**		0.234		0.495		0.131	

to those of our methods. With regard to the extra data bpn versus the watermark bpn, the watermark bpn of the CHS-based was 0.175 bpn higher than that of the NHS-based method, while the two methods required a similar quantity of extra data. The extra data required by the LS-PE-based method was 1.480 bpn more than that required by the NHS- and CHS-based methods. Chen's and Huang's methods required approximately 0.30 bpn of extra data.

The results for the test DNA sequences are summarized in Table 2, while Table 3 summarizes the results of the NHS- and CHS-based methods with $(n, k_{\max}) = (2, 2)$, Chen's method with $|w| = 2$, Huang's method with $t = 2$, and the LS-PE-based

Table 4: Occurrence probability of false codons p_f.

Occurrence probability	Proposed method		LS-PE [18]	Chen [19]	Huang et al. [20]
	NHS	CHS			
p_f	0	0	1.47×10^{-8}	8.28×10^{-4}	6.31×10^{-5}

method with $(n, p) = (2, 20), (2, 30)$. For all the test sequences, the watermark bpn of the CHS-based method was 0.147 bpn higher than that of the LS-PE-based method, 0.175 bpn higher than that of the NHS-based method, and 0.458–0.539 bpn higher than those of Chen's and Huang's methods. The extra data bpns of Chen's and Huang's methods were approximately 0.218 bpn and 0.203 bpn. However, those of the NHS- and CHS-based methods were approximately 0.315 bpn and 0.303 bpn, which are slightly higher than those of Chen's and Huang's methods. Further, those of the LS-PE-based method with $(n, p) = (2, 20)$ and $(n, p) = (2, 30)$ were 1.114 bpn and 1.788 bpn, respectively, which are higher than those of Chen's and Huang's methods. The two conventional methods require less extra data owing to their low watermark bpns. By contrast, with regard to the capacity efficiency, which is the ratio of the watermark bpn to the extra data bpn, the CHS-based method showed the highest capacity efficiency of 1.865. This means that 1 bit of extra data is required for embedding 1.865 watermark bits. Thus, the CHS-based method requires the least amount of extra data compared to the watermark.

We assume that the embedding segments of noncoding DNA do not alter regulatory gene expression and have no impact on biological function. Although our methods show a relatively high base change rate compared to the conventional methods, they do not affect coding DNA and biological function.

4.3. False Codon Occurrence. To prevent the generation of false codons in noncoding regions, we performed false codon searching of intra-/intercode values for both the embedding process and the LSB substitution of extra data. According to our results, the NHS- and CHS-based methods do not generate false codons. However, Chen's and Huang's methods and the LS-PE-based method generate false codons during the embedding process because they do not consider the constraint of the false codon.

We define the occurrence probability of false codons, p_f, as $p_f = \Pr(b_i b_{i+1} b_{i+2} =$ "ATG", "TAG", "TAA", "TGA" $\mid \mathbf{D}'^{\mathrm{nc}})$. This indicates the probability that three consecutive nucleotide bases in the watermarked noncoding sequence $\mathbf{D}'^{\mathrm{nc}}$ will become start and stop codons. We experimented 1000 times on each test sequence with different watermarks and then computed p_f. The results are summarized in Table 4. Our methods did not generate false codons. However, Huang's and Chen's methods generated false codons every 1.58 × 104 nucleotide bases and every 1.21 × 103 nucleotide bases, respectively. The LS-PE-based method, which does not consider the stop codons, generated false codons approximately every 2.13 × 107 nucleotide bases. Even though this probability is extremely low, the false codon can be fatal to biological function.

5. Conclusions

Reversible DNA data hiding can be used for repeated embedding and detection of a watermark while recovering the original DNA sequence without loss of information. Therefore, this technique can be applied to DNA storage and DNA steganography as well as to the analysis of the mutation process using an external watermark. However, most recently proposed DNA data hiding methods are irreversible. This study evaluated reversible DNA data hiding techniques using histogram shifting of noncircular and circular types while preventing biological mutation and achieving blindness and high watermark capacity.

It is extremely difficult to extend reversible image data hiding to multiple bits or to shift the histogram bin to multiple levels owing to image quality concerns. DNA data hiding has no the invisibility evaluation similar image quality with the constraints of biological function and false start/stop codons. Therefore, it is possible to extend the difference of code values of DNA sequences to multiple bits or to shift the values to multiple levels within these constraints. We coded four-character symbols of noncoding regions into integer values using the numeric order and embedded the binary watermark in two ways, namely, by shifting the NHS and CHS types to multiple levels. Next, we prevented the generation of false codons via searching of intra-/intercode values. On the basis of our experimental results, we verified that the CHS-based method has the highest watermark bpn of 0.566 bpn, which is 0.147–0.539 bpn higher than the watermark bpns of other methods, and that this method shows the highest capacity efficiency of approximately 1.865, which is 0.622–1.734 higher than that of other methods. Furthermore, we verified that false codons are not introduced by our methods, but they are introduced every 1.21 × 103–2.13 × 107 nucleotide bases by conventional methods.

Data hiding in noncoding regions may damage unknown genes or gene regulatory networks [9]. Future studies should investigate reversible DNA watermarking in coding regions by solving codon preservation and codon optimization problems with reversibility.

Conflicts of Interest

The author declares that there are no conflicts of interest.

Acknowledgments

This research was supported by the Tongmyong University of Research Grants 2015A009.

References

[1] G. M. Church, Y. Gao, and S. Kosuri, "Next-generation digital information storage in DNA," *Science*, vol. 337, no. 6102, p. 1628, 2012.

[2] N. Goldman, P. Bertone, S. Chen et al., "Towards practical, high-capacity, low-maintenance information storage in synthesized DNA," *Nature*, vol. 494, no. 7435, pp. 77–80, 2013.

[3] J. P. Cox, "Long-term data storage in DNA," *Trends in Biotechnology*, vol. 19, no. 7, pp. 247–250, 2001.

[4] M. Borda and O. Tornea, "DNA secret writing techniques," in *Proceedings of the 8th International Conference on Communications (COMM '10)*, pp. 451–456, Bucharest, Romania, June 2010.

[5] D. Tulpan, C. Regoui, G. Durand, L. Belliveau, and S. Léger, "HyDEn: a hybrid steganocryptographic approach for data encryption using randomized error-correcting DNA codes," *BioMed Research International*, vol. 2013, Article ID 634832, 11 pages, 2013.

[6] O. O. Babatunde, "Deoxyribonucleic acid (DNA) as a hypothetical information hiding medium: DNA mimics basic information security protocol," *Journal of Engineering and Technology Research*, vol. 3, no. 5, pp. 148–154, 2011.

[7] D. Heider and A. Barnekow, "DNA-based watermarks using the DNA-Crypt algorithm," *BMC Bioinformatics*, vol. 8, article 176, 2007.

[8] D. Heider, D. Kessler, and A. Barnekow, "Watermarking sexually reproducing diploid organisms," *Bioinformatics*, vol. 24, no. 17, pp. 1961-1962, 2008.

[9] D. Heider, M. Pyka, and A. Barnekow, "DNA watermarks in non-coding regulatory sequences," *BMC Research Notes*, vol. 2, article 125, 2009.

[10] D. Heider and A. Barnekow, "DNA watermarking: Challenging perspectives for biotechnological applications," *Current Bioinformatics*, vol. 6, no. 3, pp. 375–382, 2011.

[11] D. Haughton and F. Balado, "BioCode: two biologically compatible algorithms for embedding data in non-coding and coding regions of DNA," *BMC Bioinformatics*, vol. 14, no. 1, article 121, 2013.

[12] S.-H. Lee, "DWT based coding DNA watermarking for DNA copyright protection," *Information Sciences*, vol. 273, pp. 263–286, 2014.

[13] S.-H. Lee, "DNA sequence watermarking based on random circular angle," *Digital Signal Processing*, vol. 25, no. 1, pp. 173–189, 2014.

[14] I. Hafeez, A. Khan, and A. Qadir, "DNA-LCEB: a high-capacity and mutation-resistant DNA data-hiding approach by employing encryption, error correcting codes, and hybrid twofold and fourfold codon-based strategy for synonymous substitution in amino acids," *Medical & Biological Engineering & Computing*, vol. 52, no. 11, pp. 945–961, 2014.

[15] F. Balado, "On the embedding capacity of DNA strands under substitution, insertion, and deletion mutations," in *Media Forensics and Security II, 754114*, vol. 7541 of *Proceedings of SPIE*, San Jose, Calif, USA, January 2010.

[16] F. Balado, "On the Shannon capacity of DNA data embedding," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '10)*, pp. 1766–1769, March 2010.

[17] S. Lee and K. Kwon, "Consecutive difference expansion based reversible DNA watermarking," *Journal of the Institute of Electronics and Information Engineers*, vol. 52, no. 7, pp. 51–62, 2015.

[18] S. Lee, S. Kwon, and K. 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, 2015.

[19] T. Chen, "A novel biology-based reversible data hiding fusion scheme," in *Frontiers in Algorithmics*, vol. 4613 of *Lecture Notes in Computer Science*, pp. 84–95, Springer, 2007.

[20] 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, no. 3, pp. 1439–1451, 2014.

[21] J. Tian, "Reversible data embedding using a difference expansion," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 8, pp. 890–896, 2003.

[22] D. M. Thodi and J. J. Rodriguez, "Expansion embedding techniques for reversible watermarking," *IEEE Transactions on Image Processing*, vol. 16, no. 3, pp. 721–730, 2007.

[23] I.-C. Dragoi and D. Coltuc, "Local-prediction-based difference expansion reversible watermarking," *IEEE Transactions on Image Processing*, vol. 23, no. 4, pp. 1779–1790, 2014.

[24] B. Ou, X. Li, Y. Zhao, R. Ni, and Y.-Q. Shi, "Pairwise prediction-error expansion for efficient reversible data hiding," *IEEE Transactions on Image Processing*, vol. 22, no. 12, pp. 5010–5021, 2013.

[25] R. Naskar and R. S. Chakraborty, "Histogram-bin-shifting-based reversible watermarking for colour images," *IET Image Processing*, vol. 7, no. 2, pp. 99–110, 2013.

[26] G. Coatrieux, W. Pan, N. Cuppens-Boulahia, F. Cuppens, and C. Roux, "Reversible watermarking based on invariant image classification and dynamic histogram shifting," *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 1, pp. 111–120, 2013.

[27] M. U. Celik, G. Sharma, A. M. Tekalp, and E. Saber, "Lossless generalized-LSB data embedding," *IEEE Transactions on Image Processing*, vol. 14, no. 2, pp. 253–266, 2005.

[28] L.-T. Ko, J.-E. Chen, Y.-S. Shieh, H.-C. Hsin, and T.-Y. Sung, "Nested quantization index modulation for reversible watermarking and its application to healthcare information management systems," *Computational and Mathematical Methods in Medicine*, vol. 2012, Article ID 839161, 8 pages, 2012.

[29] A. Khan, A. Siddiqa, S. Munib, and S. A. Malik, "A recent survey of reversible watermarking techniques," *Information Sciences*, vol. 279, pp. 251–272, 2014.

[30] H. J. Shiu, K. L. Ng, J. F. Fang, R. C. Lee, and C. H. Huang, "Data hiding methods based upon DNA sequences," *Information Sciences*, vol. 180, no. 11, pp. 2196–2208, 2010.

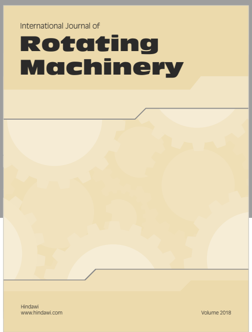
[31] G. Liu, H. Liu, and A. Kadir, "Hiding message into DNA sequence through DNA coding and chaotic maps," *Medical & Biological Engineering & Computing*, vol. 52, no. 9, pp. 741–747, 2014.

[32] J. Fu, W. Zhang, N. Yu, G. Ma, and Q. Tang, "Fast tamper location of batch DNA sequences based on reversible data hiding," in *Proceedings of the 7th International Conference on BioMedical Engineering and Informatics (BMEI '14)*, pp. 868–872, October 2014.

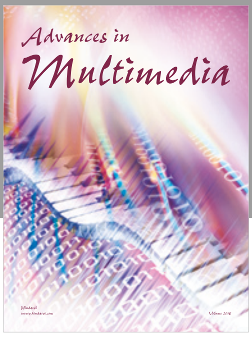
[33] G. Ma, Q. Tang, W. Zhang, and N. Yu, "Tamper restoration on DNA sequences based on reversible data hiding," in *Proceedings of the 6th International Conference on Biomedical Engineering and Informatics (BMEI '13)*, pp. 484–489, December 2013.

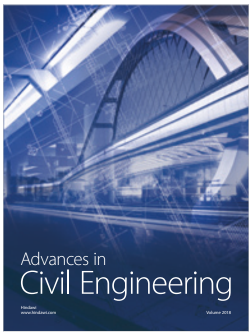
Journal of
Engineering

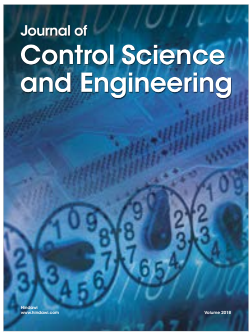
The Scientific
World Journal


International Journal of
Rotating
Machinery

Journal of
Sensors

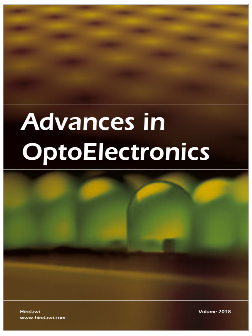

Advances in
Multimedia


Advances in
Civil Engineering


Journal of
Control Science
and Engineering

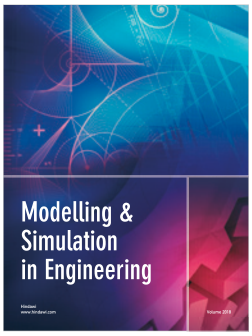
Journal of
Robotics

Journal of
Electrical and Computer
Engineering

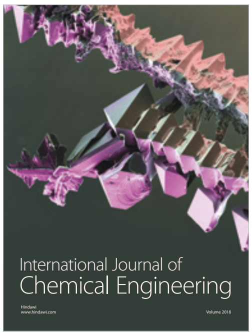

Advances in
OptoElectronics

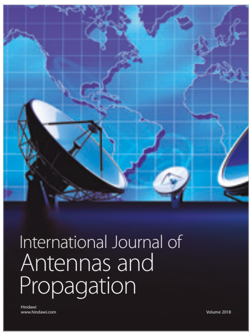
VLSI Design

International Journal of
Navigation and
Observation

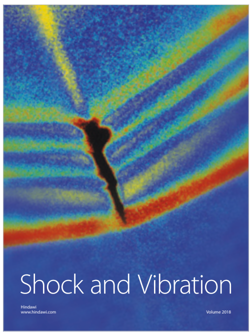

Modelling &
Simulation
in Engineering

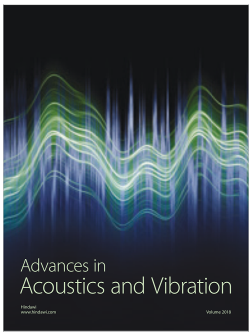
International Journal of
Aerospace
Engineering


International Journal of
Chemical Engineering


International Journal of
Antennas and
Propagation

Active and Passive
Electronic Components


Shock and Vibration


Advances in
Acoustics and Vibration