Invited Paper

A reversible watermarking for DNA sequence using an adaptive least square prediction error expansion

Suk-Hwan Lee¹a), Eung-Joo Lee², and Ki-Ryong Kwon³

¹ Department of Information Security, Tongmyong University
428, Sinseonro, Namgu, Busan, Korea

² Department of Information Communication Engineering, Tongmyong University
428, Sinseonro, Namgu, Busan, Korea

³ Department of IT Convergence and Application Engineering, Pukyong National University
45, Yongsoro, Namgu, Busan, Korea

a) skylee@tu.ac.kr, sukhwanlee@gmail.com

Received July 9, 2019; Revised September 1, 2019; Published January 1, 2020

Abstract: With the development of bio-computing technology, the research for DNA watermarking, which considers DNA information a medium, has been showing interest. In particular, there is a need for a reversible DNA watermarking technology capable of DNA storage and forgery prevention of DNA sequence, and analyzing biological mutation processes by watermark while recovering perfectly the original DNA sequence. In this paper, we address a reversible watermarking method for noncoding DNA sequences using an adaptive prediction error expansion based on least square predictor. In our method, the 4-character nucleotide sequences of the noncoding region are converted into code values by the adjacent n nucleotide bases. Then, a least squares based prediction error for the current code coefficient is obtained, and this prediction error is expanded adaptively by the number of bits determined according to the condition of prediction error expansion. Here, a false start codon generation is prevented through a comparison search between the watermarked adjacent base sequences. The experimental results showed that our method has a higher watermark capacity than the conventional method and the mean prediction error extension method, and does not generate biological mutations and false start codons.

Key Words: reversible DNA watermarking, DNA security, DNA copyright, least squares based prediction error expansion
1. Introduction

Recently, there has been a lot of research on DNA storage for large-capacity [1], DNA steganographic for secret communication [2, 3], DNA watermarking [4–8] for copyright protection in the DNA sequence. Techniques of DNA storage, DNA steganography, and DNA watermarking commonly should hide the extrinsic information in DNA sequences without altering biological function. However, DNA watermarking requires a reversible watermarking technique that can be recovered without loss of the original DNA sequence. Reversible DNA watermarking has not been studied much in comparison with irreversible DNA watermarking because of low signal intensity of 4-character sequence, unlike general multimedia data.

Of the existing reversible DNA watermarking methods, Chen et al. [9] converted 4-character strings of noncoding DNA sequence to decimal sequence and then applied a lossless compression and DE (Difference Expansion) based method, which is a typical reversible image watermarking method, in this sequence. Huang et al. [10] applied the histogram-based method for low base change rates, but have low capacity. Other methods [11–13] use the complementary paired base substitution to hide the reversible watermark but require the original sequence to be referred to when extracting and recovering watermarks. These existing methods maintain the length of original sequence but do not consider false start codon prevention [9–13], or are non-blind [11–13], or has low capacity [10]. We have been researching based on difference expansion [14, 15] and histogram shifting [16, 17] based reversible DNA watermarking to solve the false start codon prevention, non-blind, and low capacity which are the disadvantages of the conventional methods.

The main differences between reversible DNA watermarking and reversible image watermarking [18–21] are as follows: 1) Generally, similarity between neighboring code coefficients is not large in base sequence, unlike image data with many similarities between adjacent pixels. Thus, the prediction error extension [19, 20] using the similarity between neighbors, which is often used in reversible image watermarking, is not suitable for base sequence. 2) The code coefficient of base sequence can be moved within the dynamic range unless biological functions are changed or false start codons are generated. In other words, image quality is a very important consideration in reversible image watermarking. However, in reversible DNA watermarking, there are limitations such as maintaining amino acids and preventing false start codon, although movement is free within the four-character range of (A, T, C, G). Reverse image watermarking is difficult to expand the prediction error by multi-bit depending on the visual quality and extension conditions, but reversible DNA watermarking is not difficult to expand the prediction error by multi-bit according to the integer coefficient of 4-character sequence.

In this paper, we propose a reversible watermarking method for noncoding DNA sequences using an adaptive prediction error expansion based least square predictor for false start codon prevention and high capacity. Our method consists of three steps: code conversion, prediction error expansion, and false start codon prevention. First, in the code coefficient conversion process, the 4-character base sequence is converted to the integer coefficients of \( n \) consecutive base units (or \( n \) coding orders) for facilitating watermarking signal processing. At this point, a sequence of coefficients of \( 2^2n \) bits is generated according to \( n \) coding order. Second, in the prediction error expansion process, the prediction error of each code coefficient is obtained by using a least square predictor of the prediction order \( p \), and then the prediction error is expanded by the maximum number of bits that can be expanded. Then, the watermarked code coefficients and 4-character sequence are sequentially obtained by the expansion of the prediction error. In the last step, the generation of false start codon is prevented by a comparative search between the watermarked base sequence (intra mode) and the adjacent base sequence (inter mode).

Experiments have compared the bit-per-nucleotide base (bnp) and the occurrence probability of false start codon of our method with the existing Chen’s method [9], Huang’s method [10], and mean prediction error expansion method [14]. Experimental results confirmed that our method has about 1.08 times more bnp than the mean prediction error expansion method, about 3.88 times more bnp than Chen’s method, and about 15.5 times more bnp than Huang’s method, when the coding order and the prediction order are \((n, p) = (2, 2)\). Additionally, the existing methods generated a false start
codon with a probability of $1.73 \times 10^{-6} \sim 9.11 \times 10^{-5}$, which means that the noncoding region was recognized as the coding region, resulting false amino acids. However, we confirmed that our method does not cause false start codons at all.

The organization of this paper is as follows: Section 2 introduces reversible DNA watermarking and explains main existing reversible DNA watermarking methods. Section 3 explains the coding of base sequence and false start codon prevention, and LS-PPE based multi-bit expansion of the proposed method. Section 4 analyzes the performance evaluation of our method and the existing methods. Finally, we conclude this paper in the last Section 5.

2. Related works

2.1 Fundamentals of reversible DNA watermarking

DNA sequence consists of a coding region (coding DNA) and a noncoding region (noncoding DNA), and the watermark can be embedded in these two regions. Code DNA is a component of DNA encoded by a protein, and the amino acid encoded by the protein must not change by the watermark. Noncoding DNA is a component that is not encoded as a protein, and most of noncoding DNA is known as Junk DNA, which has no genetic information. However, many types of noncoding DNA sequences are known to carry out biological functions such as transcriptional and translational control of protein coding sequences.

Reversible DNA watermarking can be recovered without loss of the original DNA sequence, and can be applied to noncoding regions rather than code regions. Since noncoding DNA has no condition for preserving the protein code, the available range of watermark is somewhat higher than that of coding DNA. Noncoding DNA has no properties such as image quality of multimedia data, but there are key features to consider.

1) Integer conversion of 4-character base sequence: Since the base sequence consists of four characters (A, T, C, and G (or U)), it has a very low dynamic range compared to multimedia data. One base has 2 bits of information and has a very low level as compared with an 8-bit image pixel value. However, when consecutive bases are combined, it is possible to have multiple bits of information.

2) Prevention of false start codon generation: It is possible that some base sequences of noncoding DNA are changed to the start codon (Methionine, ‘ATG’) of the DNA sequence by the watermark. Therefore, base sequences to be changed into the start codon must be predicted in advance and excluded in the embedding process.

3) Blind detection and conservation of sequence length: Watermark detection or DNA sequence recovery should be possible with no reference sequence or original sequence without altering the length of the DNA sequence. Base sequences containing the watermark information can be appended and extracted to the junk region of the noncoding DNA sequence. However, since the identification of biological functions by addition of an external base sequence is not clear, a watermark must be embedded within a given DNA sequence.

2.2 Reversible DNA watermarking

As a representative reversible DNA watermarking method, Chen et al. [9] proposed lossless compression and DE (Difference expansion)-based methods, which are widely used in existing reversible image watermarking methods. They first convert the four-character base symbols to a 2-bit binary, then convert the binary base sequence to a decimal sequence of $|w|$ bits. And they embed secret messages into a decimal sequence based on lossless compression and difference expansion. The former lossless-based method compresses the decimal sequence by arithmetic coding, and then adds the binary secret message to the end of the compressed string. This method has the highest bpn when $|w|$ is 2 bits and has an average of 0.75 bpn $\sim$ 0.81 bpn experimentally. However, this method changes the compression coding profile and also increases the length of the compressed string. The latter DE-based method generates a position map of each word pair after classifying the word pairs of the decimal sequence into an extension set S1, a change set S2, and a non-change set S3. The method then adds the compressed location map, the original LSBs (LSB (S2)) of the word pair belonging to S2, and the secret messages to the end of the compressed string and embeds the secret message bit to the difference of
the word pairs belonging to S1. As this method applies the reversible image watermarking method of Thodi et al. [19] to the DNA sequence, it has experimentally a low data capacity of 0.09 bpn∼0.13 bpn on average when |w| is 2 bits. These two methods have the disadvantage that the false start codon is generated without considering the characteristics of noncoding DNA and the watermark data capacity is very low.

Huang et al. [9] proposed a histogram-based reversible DNA watermarking method with low base change rate. This method converts a binary base sequence into a decimal number in 2-bit units as Chen’s method, and then obtains a histogram of the decimal sequence. Here, h is the highest frequency value, L1 is the lowest frequency value, and L2 is the second lowest frequency value. If any decimal number p_i is equal to L1, then change p_i to L2 and set a map to 1. Otherwise p_i is equal to L2, then do not change p_i and set a map to 0. When p_i is h, p_i is not changed if the secret bit is 0, or p_i is changed to L1 if the secret bit is 1. DNA sequence recovery and watermark extraction are performed by the position map, h, L1, and L2 values. Experimentally, this method has 0.024 bpn and a base change rate of 4.07%∼4.80% at t = 2 and 0.011 bpn and a base change rate of 1.86%∼2.34% at t = 3. That is, this method has a low base change rate, but bpn is very low, and generate a false start codon as in Chen’s method.

In addition, Liu et al. [11] proposed a piecewise linear chaotic map (PWLCM) based watermarking method and J. Fu [12] and Ma [13] proposed a DNA reversible watermarking method for tamper location and restoration of DNA sequences. As these methods embed the watermark by the complementary rule based substitution method, it is non-blind method which requires reference (or original) DNA sequence for extraction and restoration.

3. Proposed reversible DNA watermarking method using least square predictor

The proposed processes for watermark embedding, extracting, and recovery of DNA sequences are shown in Fig. 1. A DNA sequence consists of noncoding regions and coding regions, and among noncoding regions, the embedding areas with adequate length for the watermark to be hidden are selected. The watermark embedding process consists of the integer coding of 4-character base sequence, the watermark embedding by LS prediction and expansion of the prediction error, the identification of false start codon between intra and inter-mode, and the additional information insertion by LSB substitution. Here, the watermark capacity depends on the parameters of the coding order n and the prediction order p.

3.1 Integer coding of base sequence by coding order n

A nucleotide base is represented by 4-character, b=(A, T, C, G), and each letter is represented by four decimal digits or two binary digits; b = (0, 1, 2, 3)_{10} = (00, 01, 10, 11)_{2}. We extend the 2-bit integer to an integer represented by more than 2 bits for the availability of signal processing, as shown in Fig. 2(a). Thus, we code a base block x composed of n consecutive bases as an integer x of 2n bits as follows;

\[ x = f(x) = \sum_{k=1}^{n} 2^{2(n-k)} b_k \]  

(1)

where x = (b_1, b_2, ..., b_n) and x ∈ [0, 2^{2n} - 1]. Conversely, the nucleotide bases of the base block are easily obtained from the integer value x; \( f^{-1}(x) = x \) where \( b_k = (x \gg 2(n - k)) \% 4 \) for \( k = 1, \cdots, n \). Here, \( \gg \) is the right bit-shift operator and \( \% \) is the modular operator. All nucleotide bases in the embedding region are encoded with integer values X by the coding order n.

3.2 False start codon prevention

Some bases on the nucleotide base sequence can be changed to ‘ATG’ which is the start codon of the coding region in the watermark embedding process. These false initiation codons change a portion of the noncoding region to the coding region and cause a change in the biological function. The false start codon can be generated within integer values or between integer values as follows:
1) Intra-code within integer value: If the coding order is \( n > 2 \), \( n - 2 \) false start codons can be generated within the integer value domain as shown in Fig. 2(b). Since \( 2^{2(n-3)} \) integer values including false start codons can be generated at an arbitrary position \( j \in [1, n-2] \) in a base block, there are a total of \( 2^{2(n-3)} \times (n-2) \) integer values having false start codons at \((n-2)\) positions. Our method generates an integer value table \( Z \) including a false start codon in advance, and then performs the embedding process so that the watermarked integer values are not included in \( Z \).

**Fig. 1.** Proposed reversible DNA watermarking with processes of (a) watermark embedding, (b) watermark extracting and original DNA sequence recovery.

**Fig. 2.** (a) \( 2^n \)-bit integer representation for \( n \)th base block when the coding order is \( n = 2 \) \((x \in [0, 2^{2x+4} - 1])\) and (b) Possibility of false start codon (a) within an integer value (intra-code) and between integer values (inter-code).
2) Inter-code between integer values: A false start codon may be generated between the base block $x'_{i-1}$ of the previous watermarked value $x_{i-1}$ and the base block $x'_i$ of the currently processed value $x_i$. As shown in Fig. 2(b), a false start codon occurs in the middle of (⋯,A,TG,⋯) or (⋯,AT,GT⋯) when $(x'_{i-1}, x'_i)$. Our method adjusts the number of embedding bits with respect to the integer value $x_i$ so that when the previous watermarked value $x_{i-1}$ is given, the current watermarked value $x'_i$ does not satisfy the above condition.

### 3.3 Prediction error expansion conditions for multi-bit embedding

Unlike the image data, the DNA code value without the condition for image quality is free to move within the effective range excluding the false start codon value. Therefore, the prediction error $e$ for the integer value pair can be expanded $2^k$ times for embedding $k$ bits according to the expansion condition. Given the $k$ bits of the watermark $\{w_j\}_1^k$ and the prediction value $\hat{x}$, the watermarked value $x'$ is obtained by a prediction error $e$ expanded by $2^k$ times;

$$x' = \hat{x} + 2^k e + \text{sgn}(e) \sum_{i=1}^{k} 2^{i-1} w_j$$  \hspace{1cm} (2)

where $e = x - \hat{x}$. Given the watermarked value $x'$ and the number of bits $k$, the watermark extraction and integer value recovery can be easily obtained as follows:

$$w_j = ((x' - \hat{x}) \gg (j - 1)) \% 2 \text{ for } j = 1, \cdots, k$$  \hspace{1cm} (3)

$$x = \hat{x} + e = \hat{x} + (x' - \hat{x}) \gg k$$  \hspace{1cm} (4)

Since the watermarked value $x'$ must be $0 \leq x' \leq 2^{2n-1}$, the expansion condition of the prediction error $e$ for $2^k$ times expansion is

$$e \in [2^{-k}(-\hat{x} - \text{sgn}(e) \sum_{i=1}^{k} 2^{i-1} w_j), 2^{-k}(2^2 - 1 - \hat{x} - \text{sgn}(e) \sum_{i=1}^{k} 2^{i-1} w_j)]$$  \hspace{1cm} (5)

and the value $x$ must satisfy the following condition:

$$x \in [\text{max}(0, \lceil \hat{x} + 2^{-k}(-\hat{x} - a(k)) \rceil), \text{min}(2^{2n} - 1, \lfloor \hat{x} + 2^{-k}(2^{2n} - 1 - \hat{x} - a(k)) \rfloor)]$$  \hspace{1cm} (6)

where $a(k) = \text{sgn}(e) \sum_{i=1}^{k} 2^{i-1} w_j$. Here, $\lceil \cdot \rceil$ and $\lfloor \cdot \rfloor$ are the ceiling function and flooring function. An expansion condition is determined by the watermark $\{w_j\}_1^k$ of $k$ bits and the prediction value $\hat{x}$, and the number of bits to be embedded in the integer value $x$ is determined according to the expansion condition.

Figure 3 shows the number of bits to be embedded in a value $x$ for each prediction value $\hat{x}$ when the coding order $n = 4$ and the watermark bits are all 1. Here the maximum bits for embedding is $k_{max} = 2n - 1 = 7$. The range of a value $x$ according to the number of embedded bits when the prediction value $\hat{x}$ is 0, 128, and 255, respectively. As the number of embedded bits increases, the expandable area becomes exponentially narrow. As $\hat{x}$ approaches 0 or 255, the number of bits to be embedded decreases.

### 3.4 LS-based prediction of DNA code value

In reversible image watermarking, various prediction methods were presented, including fixed predictors such as MED (median edge detector) predictor [19] used in JPEG-LS standardization, GAP (gradient-adjusted predictor) and SGAP (simplified GAP) [20] used in CALIC (context-based, adaptive, lossless image coding) and also adaptive predictors in LS (least square) predictor [21]. The adaptive LS predictor is more suitable than the fixed predictor such as MED and GAP in the DNA value sequence with a low correlation between adjacent values.

Figures 4(a), 4(b), 4(c) show integer values and their histograms of the sequence ‘AE017199’ when the coding order $n$ is 3, respectively. The integer value histogram of DNA sequence is expanded and contracted according to the coding order, but does not have a formal distribution according to the
sequence. That is, the ‘AE017199’ sequence is uniformly distributed in the remaining regions except for the four regions. In addition, the integer value sequence appears in a random form, and the correlation between adjacent values is very low. Therefore, in order to reduce the prediction error, our method predicts the integer value based on the local LS prediction [21] of Dragoi et al.

The column vectors for $p$ integer values, $x_i = (x_{i-1}, \cdots , x_{i-p})$, and $p$ parameters, $a_i = (a_1, \cdots , a_p)$ are given for predicting the current integer value $x_i$. Here, $p$ is called the predicting order. When $x_i$ is observed, the predicted value $\hat{x}_i$ of $x_i$ is defined by the linear regression function $f(a, x_i)$ as follows:

$$\hat{x}_i = f(a, x_i) = \sum_{j=1}^{p} a_j x_{i-j} = x_i' a$$  \hspace{1cm} (7)$$

Assuming that the column vector of all integer values $y = (x_1, \cdots , x_N)$ in any embedding area and the $N \times p$ matrix $X = (x_1', \cdots , x_N')$ of $N$ previous integer values are $X$, the LS prediction obtains parameters $a$

$$a = (X'X)^{-1}X'y'$$  \hspace{1cm} (8)$$

such that the square distance between $y'$ and $Xa'$ is minimized: $\|y' - Xa'\|$.

Our method predicts the integer values of DNS sequence through local prediction by embedding region rather than global prediction for whole embedding regions. Therefore, the decoding process requires $|O(n)| \times a$ additional information of the parameters $a$ by the number of embedding regions $|O(n)|$ of the DNA sequence.

A successive predictor $\hat{x}_i = x_{i-1}$ or a mean predictor $\hat{x}_i = \sum_{i=1}^{p} x_{i-j}/p$ is possible by the integer value prediction. Figure 4(d) shows the histogram of the prediction error for adjacent values, average prediction, and LS prediction for ‘AE017199’ sequence when the coding order is $n = 3, 4$. In this figure, ER represents the probability of occurrence of expanded area. The error for adjacent values has an expanded range of about 74.8% irrespective of the coding order. The mean and LS predictions have a somewhat higher ER when the coding order $n = 3$, and the higher the predicted order $p$, the higher the ER. In particular, when $n = 3$ and $p = 20$, the LS region has an expanded region of 91.6%, which is the highest. That is, when $n = 3$, it can be seen that the watermark capacity increases as the prediction order $p$ of LS becomes higher.

The prediction error histogram of the image is modeled as a Laplacian distribution, but the LS prediction error histogram of the integer value of DNA sequence is modeled as a normal distribution as in the result. $(n, p) = (3,10)$ and $(n, p) = (3,20)$ approaches to the distribution $(\mu, \sigma) = (0,20)$ and $(\mu, \sigma) = (0,19)$ respectively. Furthermore, $(n, p) = (4,10)$ and $(n, p) = (4,20)$ approaches to the distribution $(\mu, \sigma) = (0,80)$ and $(\mu, \sigma) = (0,76)$ respectively.

![Expandability condition](image)

**Fig. 3.** When the watermark bits are all 1; $w = \{1\}^{2n-1}$, the expandable area of $x$ to the prediction value $\hat{x}$.  

8
3.5 Reversible watermark embedding

When the coding order \( n \) and the prediction order \( p \) are given, our method obtains LS predictor parameters \( a \) for each embedding region, and then obtains the predicted value \( \hat{x}_i \) by LS prediction for \( i > p \) and mean prediction for \( i \leq p \).

\[
\hat{x}_i = \begin{cases} 
\sum_{j=1}^{p} a_j x_{i-j} & \text{if } i > p \\
\sum_{j=1}^{i-1} x_{i-j} & \text{if } 1 \leq i \leq p \\
0 & \text{if } i = 1
\end{cases}
\]  

(9)

After the embedding bit number \( k_i (\in [0, 2n-1]) \) is determined according to the expanded conditions of the prediction error \( e_i = x_i - \hat{x}_i \), \( k_i \) bits of the integer value \( x_i \) are embedded as follows:

\[
x_i = \hat{x}_i + 2^{k_i} e_i + \alpha(k_i), \text{where } \alpha(k_i) = \text{sgn}(e_i) \sum_{i=1}^{k_i} 2^{j-1} w_j
\]  

(10)

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

If some bases within \( x'_i \) or between \( x'_i \) and \( x'_{i-1} \) is included in the false start coding table \( Z^c \), we decrease the number of embedding bits \( k_i \) by one and then embed it by the above equation. We repeat this process until \( k_i \) is zero.

Numbers of embedded bits by integer values \( K = \{k_i\} \) and prediction parameters \( a \) by embedded regions are additional information needed to extract the watermark and recover the original sequence. The additional information shall be loaded in the watermarked area and transmitted without incurring

![Fig. 4.](image)

When the coding order is \( n = 3 \). (a) integer values, (b) histograms of integer values, and (c) histograms of difference between adjacent values of ‘AE017199’ sequences and (d) the mean error histograms of LS predicted value and mean predicted value.
false start codon and without generating any additional information. We compress losslessly \( K, a \) and \( B \), which is LSB bits of 2bit base binary, by the arithmetic coding and generate a compressed stream \( C = \{ c_i \} \). Compressed bit \( c_i \) is substituted into LSB of binary number \( b'_i \) of watermarked base as follows:
\[
b''_i = (b'_i \gg 1 + c_i, \text{ if } b'_{i-2} \neq 'A' \text{ and } b'_{i-1} \neq 'T')
\]
\( \gg \) and \( \ll \) are bitwise shift operators. If the two previously embedded bases \( (b'_{i-2}, b'_{i-1}) \) are “AT” and \( b'_i \) is ‘G’, \( b'_i \) is substituted into one of ‘A’, ‘T’ and ‘C’. If \( b'_i \) is not ‘G’, the substitution is skipped.

### 3.6 Watermark extracting and sequence recovery

In the extracting process, we firstly obtain the number of embedded bits \( K \), the prediction parameters \( a \), and the base LSB bit \( B \) of the compressed sequence of supplementary information from the LSB of all bases excluding the base following the “AT” on the noncoding region of the transmitted DNA sequence. Then, we converted the watermarked sequence in which \( B \) is substituted for the base LSB bit to integer values \( X' \) by the coding order \( n \). Finally, we extract the watermark \( w \) and recover the original integer values \( X \) by \( K \) and \( a \).

For example, given an arbitrary integer value \( x'_i \) of the number of embedded bits \( k_i > 0 \), the predicted value \( \hat{x}_i \) is obtained from the previously recovered value \( (x_{i-1, \ldots, i-p}) \) and then the watermark of \( k_i \) bits are extracted from the prediction error \( e_i = x_i - \hat{x}_i \) by \( w_i = ((x_i - \hat{x}_i) \gg (l-1)) \% 2 \) for \( l = 1, \ldots, k_i \). Then, an original integer value \( x_i \) is restored as \( x_i = \hat{x}_i + ((x'_i - \hat{x}_i) \gg k_i) \) by shifting the prediction error \( e_i \) by \( k_i \) bits.

### 4. Experimental results

#### 4.1 Experimental environment

For the performance evaluation, PSNR for the capacity (bpp; bit per pixel) is used for reversible image watermarking, but biological function change is used for reversible DNA watermarking instead of PSNR image quality measure. The proposed method embeds the reversible watermark so there is no change in biological function so that no false start codon is generated in the noncoding region. Thus, in this experiment, we compared and analyzed the watermark capacity \( bpn_W \), compressed additional information amount \( bpn_{\text{Extra}} \), base change rate \( e \), and occurrence probability of false start codon of the proposed least square prediction error expansion (LS-PE) method, the conventional Chen’s method [9], Huang’s method [10], the mean prediction expansion (Mean-PE) method [14]. Here, the base change rate \( e \) represents the ratio of base changed by watermark. Assuming that the probability that the base is changed by an arbitrary watermark has a uniform distribution as a whole, the base change rate is close to \( 3/4 = 0.75 \).

The watermark capacity and the additional information quantity of the proposed method are determined by the coding order \( n \) and the prediction order \( p \). Therefore, we have experimentally selected the prediction order of LS-PE method with additional information amount and maximum watermark capacity and then compared the capacity of each method by the coding order. For fair performance evaluation, we used the variables with the highest watermark capacity in Chen’s method \((|w| = 2)\) and Huang’s method \((t = 2)\), respectively.

The DNA sequences used in the experiments were provided by NCBI GenBank, and their types, access no., number of bases, and number of noncoding DNA regions are shown in Table I. Noncoding regions have various base numbers, and regions with very small number of bases are excluded from the embedding target.

#### 4.2 Watermark capacity and additional information amount of proposed method

The watermark capacity of the proposed method is affected by the coding order \( n \) and the prediction order \( p \). When \( n \) and \( p \) are given, the number of watermark bits embedded in the embedded regions corresponds to the sum of the number of embedded bits in each region. When \( M \) is the number of embedded regions and \( n_i \) represents the number of integer values in ith the noncoding region \( D_i \), the number of bits per base \( bpn_W(n, p) \) can be computed as follows;
Table I. DNA sequences that was tested in experiments.

| Type (Access No.) | Total bases | No. of noncoding regions | No. of bases in noncoding region |
|-------------------|-------------|--------------------------|---------------------------------|
| Archaea (AE017199) | 490,885     | 289                      | 38,932                          |
| Bacterium (CP000108) | 2,572,079   | 1,770                     | 301,761                         |
| Bacterium (CP000473.1) | 9,965,640   | 6,224                     | 962,527                         |
| Eukaryota (NC006033) | 1,195,132   | 533                       | 393,739                         |
| Moss (AP005672.1) | 122,890     | 99                        | 51,916                          |
| Plant (NC025652.1) | 141,255     | 68                        | 91,971                          |
| Virus (AY653733.1) | 1,181,404   | 883                       | 155,805                         |

\[
bpnW(n, p) = \frac{1}{M} \sum_{i=1}^{M} \left( \frac{N_i}{|D_i|} \sum_{j=1}^{N_i} k_j \right) \text{[bit/base]} \quad (12)
\]

where \(N_i = \lceil |D_i|/n \rceil\) and \(0 \leq k_j \leq 2n - 1\) and \(|D_i|\) is the cardinality of base sequence \(D_i\). Here, \(\lceil x \rceil\) is the ceiling function that maps \(x\) to the least integer greater than or equal to \(x\).

The LSB substitutable bits \(NS\) for embedding the compressed stream of additional information \(C\) is determined by the number of bases omitted by the false start codon. The maximum substitutable bits is the same as the total number of bases; \(\max(\text{NS}) = 1/M \sum_{i=1}^{M} |D_i|\). Since the length of \(C\) must be smaller than the substitutable bits \(\text{NS}\), the additional information of the number of embedded bits \(K\), the prediction parameters \(a\), and LSBs of binary bases \(B\) are small, or an algorithm with high compression efficiency is needed.

Given an arbitrary watermarked region \(D_i\), \(B\) consists of \(|D_i|\) bits, \(K\) is represented by \(N_i \lceil \log_2 2n \rceil\) bits, and \(a\) for each embedded region is represented by 32-bit floating point of \(p\) prediction orders. Therefore, the additional information \(\text{Extra}_{\text{LS–PE}}(n, p) = \sum_{i=1}^{M} (N_i \lceil \log_2 2n \rceil + |D_i| + 32p)\) [bit]. When the compression ratio of the compressed stream of additional information \(C\) is \(\rho\), the number of bits of additional information per base \(bpn_{\text{Extra}}(n, p)\) is defined as follows:

\[
bpn_{\text{Extra}}(n, p) = \frac{1}{M} \sum_{i=1}^{M} \rho(N_i \lceil \log_2 2n \rceil + |D_i| + 32p) \text{[bit/base]} \quad (13)
\]

When the coding order \(n\) is given, the larger the prediction order \(p\), the higher the occurrence probability of expansion. Figure 5(a) shows the watermark \(bpn\) \(bpn_W\) of the LS-PE method and average prediction method when \(n\) is given as 2, 3, and 4. It can be seen that \(bpn_W\) becomes larger as \(n\) is smaller and \(p\) is larger. Also, it can be seen that \(bpn_W\) of the average prediction method is substantially constant when \(p\) is 3. According to \(n\) and \(p\), \(bpn_W\) of LS-PE method is about 0.02 bpn higher than \(bpn_W\) of average prediction method. When \(n\) is 2, \(bpn_W\) of LS-PE method increases slightly from 0.413 to 0.419 since \(p\) is 20.

Fig. 5. (a) Average watermark bpn of all test sequences for coding order \(n(\in [2, 10])\) and prediction order \(p(\in [2, 30])\), (b) base change rate versus watermark bpn of LS-PE method when the coding order \(n\) are 2, 3, and 4.
The larger the prediction order $p$, the larger the watermark bpn as well as the additional information amount and base change rate. Figure 8 shows the watermark bpn versus the base change rate and also the watermark bpn versus the compressed additional information amount of the LS-PE method when the coding order $n$ is 2, 3, and 4, respectively. As shown in Fig. 5(b), it can be seen that the lower the coding order, the lower the base change rate and the higher the watermark bpn. Here, the base change rate of 0.5 indicates that 50% of the bases to be embedded are changed to one of the other three bases.

When $(n, p) = (2, 20)$, watermark bpn is 0.414, additional information bpn is 0.854, base change rate is 0.548. And, when $(n, p) = (2, 30)$, watermark bpn is 0.419, additional information bpn is 0.988, base change rate is 0.599. In our experiment, we set the prediction order $p$ with high watermark bpn to 20 while considering the additional information amount and base change rate, and analyzed the performance comparison with other HS and CHS methods.

4.3 Watermark bpn, additional information bpn, base change rate

In this experiment, we set the prediction order $p$ to 20, 30, and evaluated the watermark bpn $bpn_W$, the additional information amount bpn $bpn_{Extra}$, the base change rate, and capacity efficiency $bpn_W/bpn_{Extra}$ of the proposed LS-PE method, mean-PE method, and conventional methods varying the coding order $n$ from 2 to 6. Since conventional methods are not related to the coding order $n$, they have a single result. The results are shown in Fig. 6.

As shown in Fig. 6(a), $bpn_W$ of LS-PE method and mean-PE method decreases as the coding order

![Fig. 6. (a) Watermark bpn, (b) additional information bpn, (c) base change rate, and (d) capacity efficiency (ratio of watermark bpn and additional information bpn) of proposed LS-PE method and mean-PE method ($p = 20, 30$), conventional methods of Chen and Huang that are not depend on the coding order $n$.](image-url)
When \( n \) increases, \( \text{bpm}_{W} \) of LS-PE method is 0.419 bpn when \((n, p) = (2,30)\) and 0.413 bpn when \((n, p) = (2,20)\). Next, \( \text{bpm}_{W} \) of mean-PE method is 0.386 bpn and 0.377 bpn when \((n, p) = (2,30)\) and \((n, p) = (2,20)\), respectively. When \((n, p)\) is varied, it can be seen that LS-PE method is from 0.022 bpn to 0.049 bpn higher than mean-PE method. The watermark bpn of Chen’s method and Huang’s method was very low, with 0.109 bpn and 0.026 bpn respectively, regardless of the coding order.

As shown in Fig. 6(b), the result of the additional information bpn required for watermark extraction, \( \text{bpm}_{\text{Extra}} \) of LS-PE method was high at 0.908 bpn and 0.854 bpn, respectively, at \((n, p) = (2,30), (2,20)\). When \((n, p)\) is varied, LS-PE method has \( \text{bpm}_{\text{Extra}} \) higher than Mean-PE method by 0.044 bpn \( \sim \) 0.125 bpn. This indicates that the more watermark bpn is, the more additional data for extraction is needed. Since Chen and Huang methods have low watermark bpn, low additional data of 0.302 bpn and 0.289 bpn are needed.

As shown in Fig. 6(c), the capacity efficiency \( \text{bpm}_{W}/\text{bpm}_{\text{Extra}} \) of LS-PE method and mean-PE method is about 0.371 to 0.485, which indicates that a watermark of 0.371 to 0.485 bits per one bit of additional data can be embedded. However, Chen and Huang methods have low capacity efficiencies of 0.361 and 0.090.

As shown in Fig. 6(d), which is the result of base change rate by watermark, the higher the watermark bpn, the higher the base change rate. That is, the base change rate of the LS-PE method is as high as 0.599 and 0.549 when \((n, p)\) is \((2,30)\) and \((2,20)\) and close to 0.40 when \(n > 2\). The base change rate of mean-PE method is about 0.45 when \(n = 2\) and close to 0.38 when \(n > 2\). However, Chen and Huang methods have very low base change rates of 0.171 and 0.044 with low watermark bpn.

### 4.4 Occurrence probability of false start codon

The probability \( p_f \) that an arbitrary ternary bases becomes “ATG” in the watermarked noncoding region is called an occurrence probability of false start codon. We computed the probability \( p_f \) by testing 1,000 times repeatedly for all test DNA sequences. The proposed method performs a comparison search process to prevent false start codon in watermark embedding process and LSB substitution process of additional information data. Therefore, we confirmed that the proposed LS-PE method and mean-PE method do not generate false start codon in all experiments. However, Chen and Huang methods do not consider false start codon generation, and false start codon occurred in our experiments. That is, we have found that Chen’s method generates one false start codon per \( 10^4 \) bases and Huang’s method generates one false start codon per \( 5.78 \times 10^5 \) bases.

### 5. Conclusions

Reversible DNA watermarking can be performed repeatedly watermark embedding and extracting while recovering the original DNA without loss. Therefore, this technique enables DNA storage, steganography as well as mutation process investigation by external watermark. Since most of the DNA watermarking methods are irreversible, reversible DNA watermarking techniques with no biological variation, blind detection, and high watermark capacity are needed. In this paper, we propose a reversible DNA watermarking method based on an adaptive LS-PE (least squares prediction error expansion) using noncoding DNA regions. The proposed method enables multiple-bit embedding on the basis of base sequence coding by code order and least squares error extension by prediction order and prevents the generation of false start codon, which can be recognized as the coding DNA by the searching false start codon in intra-code and inter-code. From the experimental results, we confirmed that LS-PE method has a watermark bpn of about 0.033 bpn \( \sim \) 0.395 bpn higher than mean-PE method, Chen’s method, and Huang’s method when the coding order is 2 and the prediction order is 30 and also the capacity efficiency of LS-PE method is about 0.46, which is higher than the conventional method by about 0.101 \( \sim \) 0.372. We also found that the false start codon is not generated by LS-PE method and mean-PE method but it is generated by the conventional methods with a probability of \( 10^{-4} \sim 5.78 \times 10^{-5} \).
Acknowledgments

This research is supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT (No.2016R1D1A3B03931003 and No.2017R1A2B2012456).

References

[1] N. Goldman, P. Bertone, S. Chen, C. Dessimoz, E.M. LeProust, B. Sipos, and E. Birney, “Towards practical high-capacity, low-maintenance information storage in synthesized DNA,” *Nature*, vol. 494, pp. 77–80, February 2013.
[2] C.T. Clelland, V. Risca, and C. Bancroft, “Hiding messages in DNA microdots,” *Nature*, vol. 399, pp. 533–534, June 1999.
[3] S.-H. Lee and K.-R. Kwon, “DNA information hiding method for DNA data storage,” *Journal of The Institute of Electronics and Information Engineers*, vol. 51, no. 10, pp. 118–127, October 2014.
[4] D. Heider and A. Barnekow, “DNA-based watermarks using the DNA-Crypt algorithm,” *BMC Bioinformatics*, vol. 8, no. 176, May 2007.
[5] D. Heider and A. Barnekow, “DNA watermarks - A proof of concept,” *BMC Bioinformatics*, vol. 9, no. 40, April 2008.
[6] M. Liss, D. Daubert, K. Brunner, K. Kliche, U. Hammes, A. Leierher, and R. Wagner, “Embedding permanent watermarks in synthetic genes,” *PLOS ONE*, vol. 7, no. 8, e42465, August 2012.
[7] S.-H. Lee, “DWT based coding DNA watermarking for DNA copyright protection,” *Information Sciences*, vol. 273, pp. 263–286, July, 2014.
[8] S.-H. Lee, “DNA sequence watermarking based on random circular angle,” *Digital Signal Processing*, vol. 25, pp. 173–189, February 2014.
[9] T. Chen, “A novel biology-based reversible data hiding fusion scheme,” *Frontiers in Algorithmics, Lecture Notes in Computer Science*, vol. 4613, pp. 84–95, 2007.
[10] 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, June 2014.
[11] G. Liu, H. Liu, and A. Kadir, “Hiding message into DNA sequence through DNA coding and chaotic map,” *Medical & Biological Engineering & Computing*, vol. 52, no. 9, pp. 741–747, September 2014.
[12] J. Fu, W. Zhang, N. Yu, G. Ma, and Q. Tang, “Fast tamper location of batch DNA sequences based on reversible data hiding,” *7th International Conference on Biomedical Engineering and Informatics (BMEI)*, pp. 868–872, October 14–16, 2014.
[13] G. Ma, Q. Tang, W. Zhang, and N. Yu, “Tamper restoration on DNA sequences based on reversible data hiding,” *6th International Conference on Biomedical Engineering and Informatics (BMEI)*, pp. 484–489, December 16–18, 2013.
[14] S.-H. Lee and K.-R. Kwon, “Consecutive difference expansion based reversible DNA watermarking,” *Journal of The Institute of Electronics and Information Engineers*, vol. 52, no. 5, pp. 51–62, July 2015.
[15] S.-H. Lee, S.-G. Kwon, and K.-R. 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, November 2015.
[16] S.-H. Lee, “Reversible data hiding for DNA sequence using multilevel histogram shifting,” *Security and Communication Networks*, vol. 2018, Article ID 3530969, March 2018.
[17] S.-H. Lee, E.-J. Lee, W.-J. Hwang, and K.-R. Kwon, “Reversible DNA data hiding using multiple difference expansions for DNA authentication and storage,” *Multimedia Tools and Applications*, vol. 77, no. 15, pp. 19499–19526, August 2018.
[18] J. Tian, “Reversible watermarking using a difference expansion,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 8, pp. 890–896, August 2003.

[19] D.M. Thodi et al. “Expansion embedding techniques for reversible watermarking,” *IEEE Trans. Image Processing*, vol. 16, no. 3, March 2007.

[20] D. Coltuc, “Improved embedding for prediction-based reversible watermarking,” *IEEE Trans. Information Forensics and Security*, vol. 6, no. 3, pp. 873–882, September 2011.

[21] I.-C. Dragoi and D. Coltuc, “Local-prediction-based difference expansion reversible watermarking,” *IEEE Trans. Image Processing*, vol. 23, no. 4, pp. 1779–1790, April 2014.