REVERSIBLE DNA INFORMATION HIDING

1SUK-HWAN LEE, 2EUNG-JOO LEE, 3KI-RYONG KWON

1Dept. of Information Security, Tongmyong University, South Korea
2Dept. of Communication Engineering, Tongmyong University, South Korea
3Dept. of IT Convergence and Application Engineering, Pukyong National University, South Korea
Email: Korcaskylee@tu.ac.kr, ejlee@tu.ac.kr, krkwon@pknu.ac.kr

Abstract- In this paper, we address two ways of difference expansion based reversible DNA watermarking using noncoding DNA sequence. Unlike image data, the reversible DNA watermarking should consider the string structure of a DNA sequence, the organism functionality, the perfect recovery, and the high embedding capacity. We convert the string sequence of four characters in noncoding region to the decimal coded values and embed the watermark bit into coded values by using pairs of neighbor coded values and consecutive DE-MBE (C-DE-MBE) using the previous embedded coded values as the current estimated one. Using sequences of bacterium and archaea types, experimental results verified that our approaches have more high embedding capacity than conventional methods and produce no false start codon and recover perfectly the host sequence without the reference sequence. Especially C-DE-MBE can embed more high two times than DE-MBE.

Index Terms- Reversible DNA information hiding, DNA storage, DNA steganography, DNA watermarking, Difference Expansion

I. INTRODUCTION

Most methods for DNA storage, DNA steganography, andDNA watermarking are irreversible that the host DNAsequence cannot be recovered from the watermarked DNasequence. The reversibility is positively necessary for DNAinformation hiding without changing the biological function. Reversible DNA watermarking has been worked by someresearchers. Chen et al [1] transformed 4-character stringof noncoding DNA sequence to decimal numbers and applied them to loss less compression and DE(difference expansion). Huang et al. [2] shifted the lowest and highestestimogram depending on the watermark to low modification rate. However, this method has low data capacity. Shiu et al. [3] and Ma et al. [4] used the complementary base pairingrule for the base substitution, which they need thereference sequence for detecting and recovering. Conventional methods preserve the sequence length but do not consider the false start codon, non-blind, or havelow data capacity.

DNA sequences have been used as media for informationcommunication and storage similar to audio, image, video, and 3D multimedia data. Four-character strings of nucleotide bases consist of coding and noncoding DNA. Tostore and transmit non-genetic information, noncodingDNA is a better target than coding DNA because it does not encod proteins. However, some noncoding DNA regions function as genetic switches that regulate gene expression and many additional noncoding regions have unknown functions. The non-genetic informationshould therefore be embedded in the noncoding DNA ofnon-living or primitive organisms. Figure 1 illustrates the general process of reversible watermarking for noncodingDNA. Nucleotide bases in noncoding sequences are codedas values in a decimal or binary format and the error-corrected or encrypted watermark is embedded into these values using a reversible algorithm. Then, the watermarked values are inversely transformed tonucleotide bases of noncoding sequences and are combined with coding sequences to generate the final watermarked DNA sequence.

This paper address two ways for DE (difference expansion) based reversible DNA watermarking method using noncoding DNA with the no false start codon, theblind detection and recovery, and the high watermark capacity. Reversible watermarking for multimedia data has been much studied for long time. DE and PE (prediction error expansion) are the most popular theory for reversible image watermarking. There are two main differences between image data and DNA nucleotide base in viewpoint of processing. The first is that the similarity ofneighbor bases in DNA sequence is less, which makes it difficult to apply PE that uses the high similarity of neighbor pixels in image. The second is that DNA sequence has the biological limitation such as false start codon, codon optimization, and protein preservation, unlike image quality such as PSNR. Nucleotide bases in noncoding DNA can be freely changed under the biological limitation. Therefore, multiple bits can be embedded into each code value according to maximum expansion condition. Experiments indicated that watermark
II. PROPOSED REVERSIBLE DNA INFORMATION HIDING METHOD

We describe a DE-based reversible DNA information hiding using noncoding DNA with a high capacity, the preservation of biological function, and blindness. Our method has three steps: selection of target noncoding sequence, numeric coding of 4-character strings, and DE-based reversible watermarking. The first two steps are preprocessing for DE-based reversible watermarking. Let us look at the preprocessing steps and two methods for DE-based reversible watermarking.

A. Numeric Coding of Nucleotide bases

Typically a quaternary symbol (4-character) of nucleotide base, \( b = \{ A', T', C', 'G' \} \), can be coded to a decimal or binary value with low range. The range of code values needs to be expanded for the watermarking availability. We code nucleotide bases, denoted as \( x \), to a code value \( x \) of \( 2n \) bit.

\[
x = f(x) = b_1 \cdot 2^{2(n-1)} + \cdots + b_n \cdot 2^2.
\]

Each nucleotide base is recovered from a code value \( x \);

\[
f^{-1}(x) = x = (b_1, b_2, \ldots, b_n)
\]

where \( b_j = \left( x \gg 2(n - j) \right) \% 2 \).

\( \gg \) is the right bit-shift operator. The number \( n \) of nucleotide bases of \( x \) is called the numeric order.

B. Multiple Bit Embedding Expandability

Given \( k \) bits of watermark, \( [w_j]_k \), a code value \( x \), and a reference value \( \hat{x} \), the difference \( d = x - \hat{x} \) expands to \( 2^k \) times while including \( k \) bits. Hence, a watermarked code value \( x' \) can be obtained from the following equation;

\[
x' = \hat{x} + 2^k d + \text{sgn}(d) \sum_{j=1}^{k} 2^{j-1}w_j.
\]

With the knowledge of the number \( k \) of embedded bits, the decoding process extracts a watermark \( [w_j]_k \) and recovers a code value \( x' \) from \( x' \). Bits of watermark \( [w_j]_k \) and a reference code value \( \hat{x} \) determine the expandable condition of \( d \). Given \( \hat{x} \), a code value \( x \) should be within the available range \( R_k(x, \hat{x}, [w_j]_k) \) for \( d \) to expand \( 2^k \) times.

Figure 2 illustrates the available range of code value \( x \) for embeddable bits on a reference code value \( \hat{x} \). The available region narrows exponentially depending on embeddable bits. The available region over 1bit has the same size of the unavailable region.

C. Consecutive DE-MBE Method

1) Encoding step

-C-DE-MBE method uses a reference code value \( \hat{x} \), as a previous embedded code value \( x_{i-1} \) consecutively to more improve the watermark capacity of DE-MBE.

Given a pair of code values \( (x_i, x_{i-1}) \) and a difference of them \( d_i = x_i - x_{i-1} \) \((x_0 = 0)\), the encoding step determines the embeddable bit number \( k_i \) by the available range of \( x_i \) in Equation (12) and embeds \( k_i \) bits \( [w_j]_k \) by expanding the difference.

\[
\begin{align*}
2) x'_i &= x_{i-1} + 2^k d_i + \alpha(k_i) \\
\text{whereas} \alpha(k_i) &= \text{sgn}(d_i) \sum_{j=1}^{k_i} 2^{j-1}w_j 
\end{align*}
\]

Similarly DE-MBE method, if nucleotide bases \( (x'_{i-1}, x_{i-1}) \) of \( (x'_{i-1}, x_i) \) include the false start codon, we decrease \( k_i \) bits by one bit and embeds it into the difference. This process repeats until \((x'_{i-1}, x_i)\) do not include the false start codon or \( k_i \) is zero. After processing all pairs, we obtain finally a watermarked noncoding region \( D^{nc}_k \cdot k_i=0 \) indicates that the code value do not satisfy the expandable condition or include the false start codon. C-DE-MBE method uses (3\# all code values with \( k_i \geq 0 \) for embedding.
The overhead for extracting and recovering process is the embedded bit number for all pairs, \( K = \{k_i\}_{i=1}^{N_k} \), and LSBs of all nucleotide bases of \( D_{nk}^{\text{nc}} \), \( B = \{b_j\}_{j=1}^{\lfloor |D_{nk}^{\text{nc}}|/2 \rfloor} \) where \( N_k = \lfloor |D_{nk}^{\text{nc}}|/2 \rfloor \). Hence, the overhead is \( N_k \log_2 n + \lfloor |D_{nk}^{\text{nc}}| \rfloor \) bits. We embed the compressed stream \( C \) of overhead into LSBs of binary nucleotide bases \( \{b_j\} \) in \( D_{nk}^{\text{nc}} \) by the substitution and then finally obtain a watermarked noncoding region including overhead \( D_{nk}^{\text{w}} \). Similarly DE-MBE method, we skip nucleotide base \( b_j \) in substitution if \( (b_{j-1}, b_{j-2}) \) is “AT” or change \( b_j \) to “G” if \( (b_{j-1}, b_{j-2}, b_j) \) is “ATG”. The compress ratio \( \rho \) should satisfy the following condition enough to embed \( \sin D_{nk}^{\text{nc}} \).

\[
\rho < \frac{\lfloor |D_{nk}^{\text{nc}}| \rfloor}{N_k \log_2 n + \lfloor |D_{nk}^{\text{nc}}| \rfloor} < \frac{1}{2 \log_2 n + 1}
\]

2) Decoding step: Similarly DE-MBE method, the decoding step gets the embedded bit number \( K \) and LSBs \( B \) of binary nucleotide bases of \( D_{nk}^{\text{nc}} \) from the compressed stream \( C \) in \( D_{nk}^{\text{nc}} \) and then obtains \( D_{nk}^{\text{nc}} \) by substituting \( B \) to LSBs of binary nucleotide bases of \( D_{nk}^{\text{nc}} \). The decoding step extracts bits of watermark and recovers a host noncoding region \( D_{nk}^{\text{nc}} \) by using \( K \). Thus, given a pair of \( (x_{j-1}, x_j) \) and an embedded bit number \( k_i \), bits of watermark \( \{w_j\}_{j=1}^{k_i} \) and a code value \( x_j \) can be extracted and recovered respectively as follows.

\[
w_j = ((x_j - x_{j-1}) \gg (j-1)) \% 2 \quad \text{for } j = 1, \ldots, k_i
\]

\[
x_j = x_{j-1} + ((x_j - x_{j-1}) \gg k_i)
\]

3) Watermark capacity: The embedded bits of watermark \( W_k^{\text{CDE}} \) in a noncoding region \( D_{nk}^{\text{nc}} \) is the sum of embedded bit number on each pair \( K = \{k_i\}_{i=1}^{N_k} \). Hence bit per nucleotide base \( \text{bpm}_{k}^{\text{CDE}} \) is as follows.

\[
\text{bpm}_{k}^{\text{CDE}} = \frac{\text{w}_{k}^{\text{CDE}}}{|D_{nk}^{\text{nc}}|} = \frac{1}{|D_{nk}^{\text{nc}}|} \sum_{j=1}^{N_k} k_j \quad \text{[bit/base]}
\]

C-DE-MBE method has the watermark capacity about two times more than DE-MBE method.

II. EXPERIMENTAL RESULTS

Our experiments analyzed the capacity and reversibility of our method of DE-MBE and CDE-MBE and compare them with Chen’s method [1] and Huang’s method [2]. We compared the watermark \( \text{bpm} \) and change rate of embedded bases \( e \) of our methods in \( n = 3 \), Chen’s method with \( w = 2 \) and Huang’s method \( t = 2 \), which they reported that two parameters \( w, t \) are best in their experiments. Table 1 shows results on each test sequence. \( \text{bpm} \) of C-DE-MBE is more higher than \( \text{bpm} \) of other methods; 2.16 times, 9.03 times as many as Chen’s and Huang’s \( \text{bpm} \). Furthermore, \( \text{bpm} \) of DE-MBE is 1.09 times, 4.57 times as many as Chen’s and Huang’s \( \text{bpm} \).

Our method prevents the occurrence of false start codon by comparison searching in watermark embedding process and overhead LSB substituting process. Experiments verified that no false start codons occurred in all test sequences. However, since Chen’s method and Huang’s method do not consider the false start codon, our experiment verified that they produced false start codons in the embedding process.

Given a host noncoding sequence \( \Gamma^{nc}(n) \) and a watermarked noncoding sequence \( \Gamma^{wc}(n) \), the change rate \( e(n) \) of nucleotide bases is defined as follows.

\[
e(n) = \Gamma^{nc}(n) - \Gamma^{wc}(n)
\]

\[
e(n) = \Gamma^{nc}(n) \frac{1}{|D_{nk}^{\text{nc}}|} \sum_{j=1}^{N_k} e(k_j)
\]

where \( e(k_j) = \begin{cases} 1, & \text{if } b_{k_j} \neq b_{k_i} \\ 0, & \text{if } b_{k_i} = b_{k_i} \end{cases} \)

This indicates the rate of all changed nucleotide bases by the watermark. Figure 7(b) shows \( e(n) \) for \( n = \{2, 10\} \) of our methods. \( e(n) \) depends on the number of embedded code values or nucleotide bases. (Q) DE-MBE method uses all code values for watermarking except for ones including false start codon or getting out expandable condition. However, DE-MBE method uses a half of all code values for embedding. Hence, C-DE-MBE method has double watermark \( \text{bpm} \) and double change rate \( e(n) \) than DE-MBE method. \( e(n) \) of DE-MBE and C-DE-MBE methods approach to 0.189 and 0.370 respectively starting from \( n = 3 \). The methods developed by Chen and Huang have relatively lower watermark \( \text{bpm} \) than those of our methods. For their methods, the rates of change of nucleotide bases are 0.171 and 0.044, respectively, which are lower than those of our methods, as shown in Table 1. In particular, since Huang’s method focused on watermarking with a low mutation rate, the watermark \( \text{bpm} \) rate of change are much lower than those of other methods.
algorithm and substituted them to substitutable bits $\Psi$ among LSBs of binary nucleotide bases. Figure 8 shows mean of substitutable bits $\Psi$ and compressed bitstream $C$ for test DNA sequences in case of the numeric order $n=3$. These results know the relation of $C$ and $\Psi$; $C \approx 0.42\Psi$ in DE-MBE and $C \approx 0.53\Psi$ in C-DE-MBE. Thus, compressed bitstream can be substituted twice on substitutable bits $\Psi$. Of course, these results can be improved by lossless compression algorithm.

CONCLUSIONS

This paper address an algorithm for reversible DNAwatermarking that preserves amino acids, prevents falsestart codon, extracts and recovers without referencesequence. Even though DNA watermarking needs thereversibility, this research has not progressed much, comparing to irreversible DNA watermarking or information hiding. Our C-DE-MBE method had the highest watermark bpn; 2 times of DE-MBE, 2.16 times of Chen’s method, and 9.03 times of Huang’s method.

ACKNOWLEDGEMENT

This work was supported by 2013 Brain Busan (BB21) Project and Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (NRF-2014R1A1A4A01006663, and NRF-2016R1A1A3B03931003).

REFERENCES