US 20190251268A1

(19) United States

(12) Patent Application Publication

(10) Pub. No.: US 2019/0251268 A1

(43) Pub. Date: Aug. 15, 2019

Lee et al.

(54) REVERSIBLE DNA INFORMATION HIDING METHOD BASED ON PREDICTION-ERROR EXPANSION AND HISTROGRAM SHIFTING

(71) Applicant: **Tongmyong University Industry-Academy Cooperation Foundation**, Busan (KR)

(72) Inventors: **Sukhwan Lee**, Gimhae (KR); **Eungju Lee**, Busan (KR); **Dong Yeop Lee**, Busan (KR); **Ju Hyeon Jeong**, Busan (KR)

(21) Appl. No.: **15/905,121**

(22) Filed: **Feb. 26, 2018**

(30) **Foreign Application Priority Data**

Feb. 13, 2018 (KR) 10-2018-017337

Publication Classification

(51) **Int. Cl.**

G06F 21/60	(2006.01)
G06F 19/28	(2006.01)
G06N 3/12	(2006.01)

(52) **U.S. Cl.**

CPC ***G06F 21/60*** (2013.01); ***G06N 3/123*** (2013.01); ***G06F 19/28*** (2013.01)

(57) **ABSTRACT**

Disclosed is a reversible DNA information hiding method based on prediction-error expansion and histogram shifting, the method being capable of false start codon prevention, original sequence length preservation, high watermark capacity, and blind detection based on prediction-error expansion and histogram shifting without biological mutation.

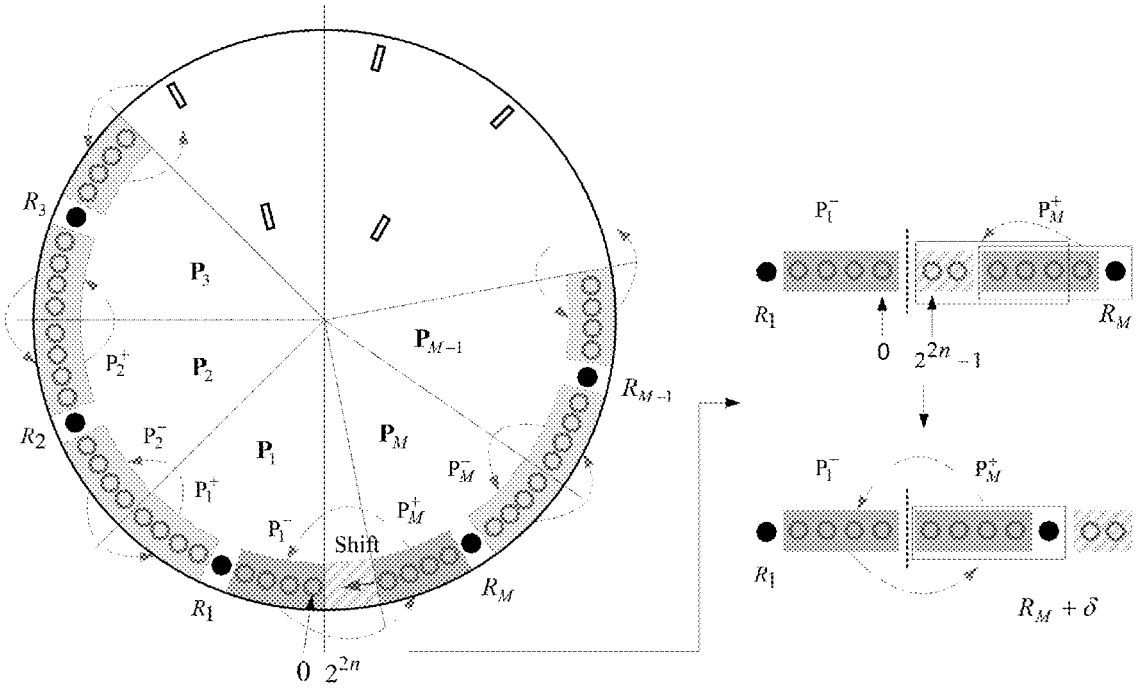


FIG. 1A

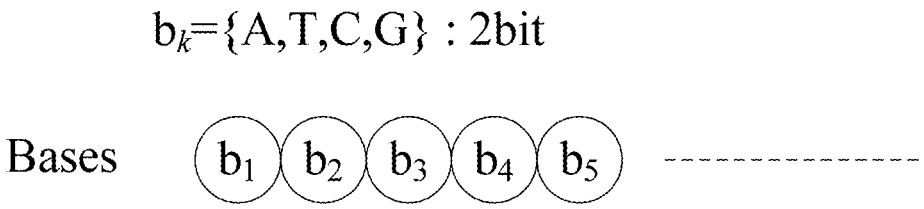


FIG. 1B

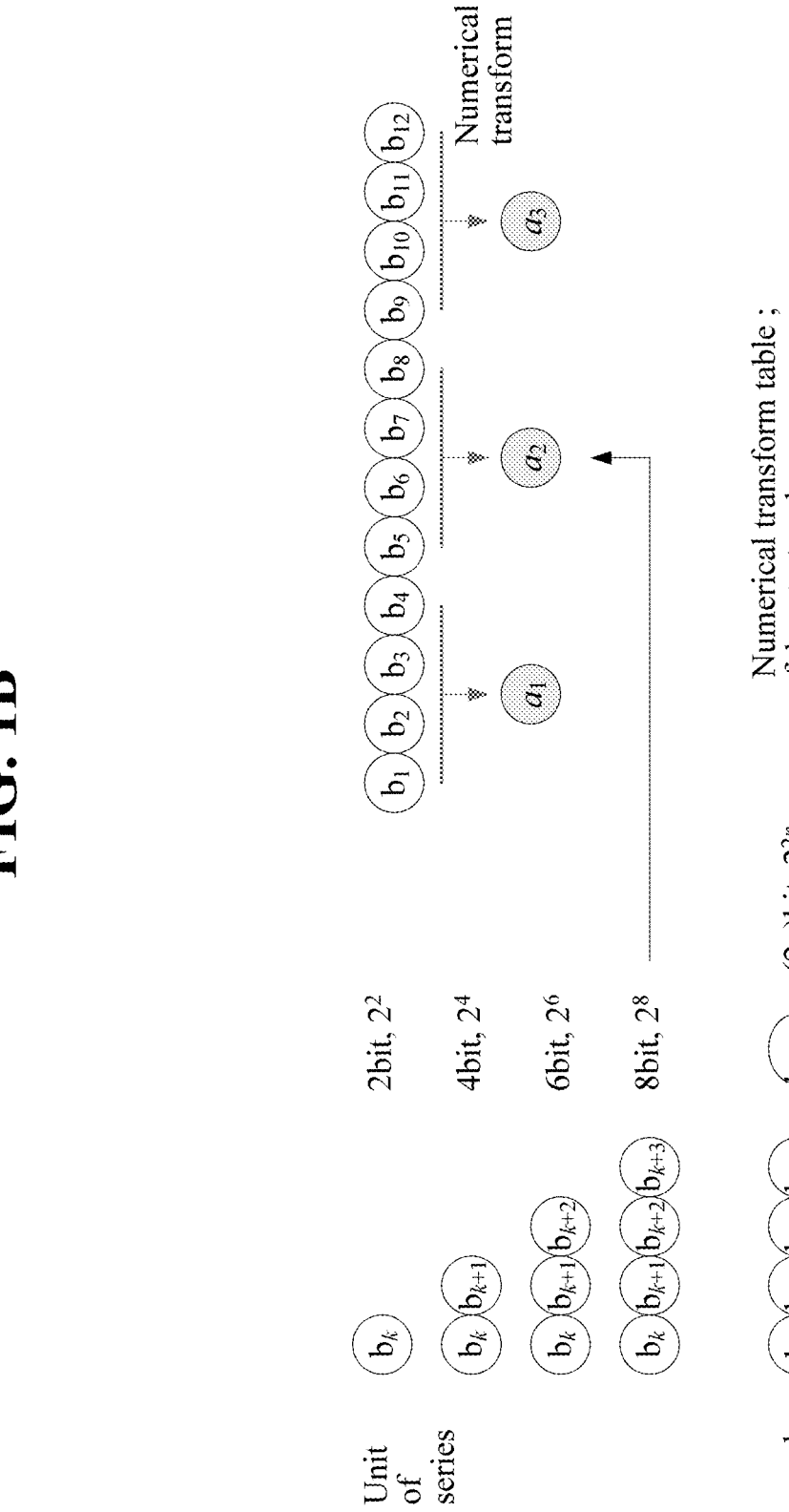
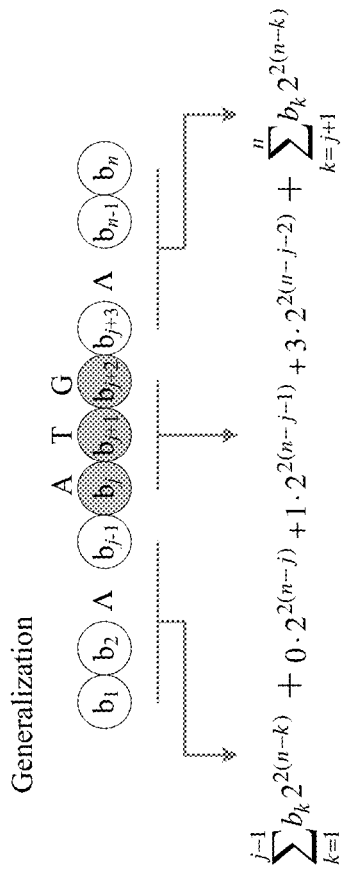

Unit of series
2bit, 2^2
4bit, 2^4
6bit, 2^6
8bit, 2^8
$(2n)$bit, 2^{2n}
n order
b_k
b_{k+1}
b_{k+2}
b_{k+3}
b_{k+n-1}
b_1 b_2 b_3 b_4 b_5 b_6 b_7 b_8 b_9 b_{10} b_{11} b_{12}
a_1
a_2
a_3
Numerical transform
Numerical transform table ;
false start codon

FIG. 2A


Generalization
A T G
b_1 b_2 Λ b_{j-1} b_j b_{j+1} b_{j+2} b_{j+3} Λ b_{n-1} b_n
$\sum_{k=1}^{j-1} b_k 2^{2(n-k)} + 0 \cdot 2^{2(n-j)} + 1 \cdot 2^{2(n-j-1)} + 3 \cdot 2^{2(n-j-2)} + \sum_{k=j+1}^{n} b_k 2^{2(n-k)}$

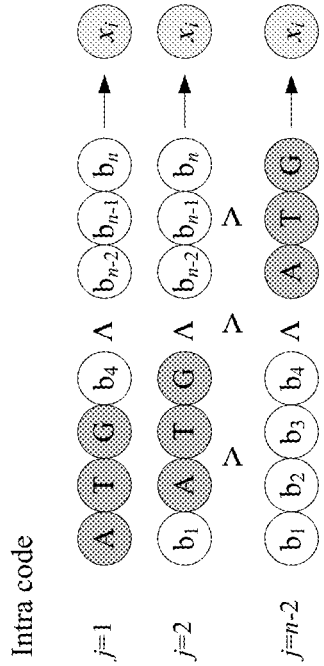

Intra code
j=1
A T G b_4 Λ b_{n-2} b_{n-1} b_n x_1
j=2
b_1 A T G Λ b_{n-2} b_{n-1} b_n x_2
Λ
Λ
j=n-2
b_1 b_2 b_3 b_4 Λ A T G x_{n-2}

FIG. 2B

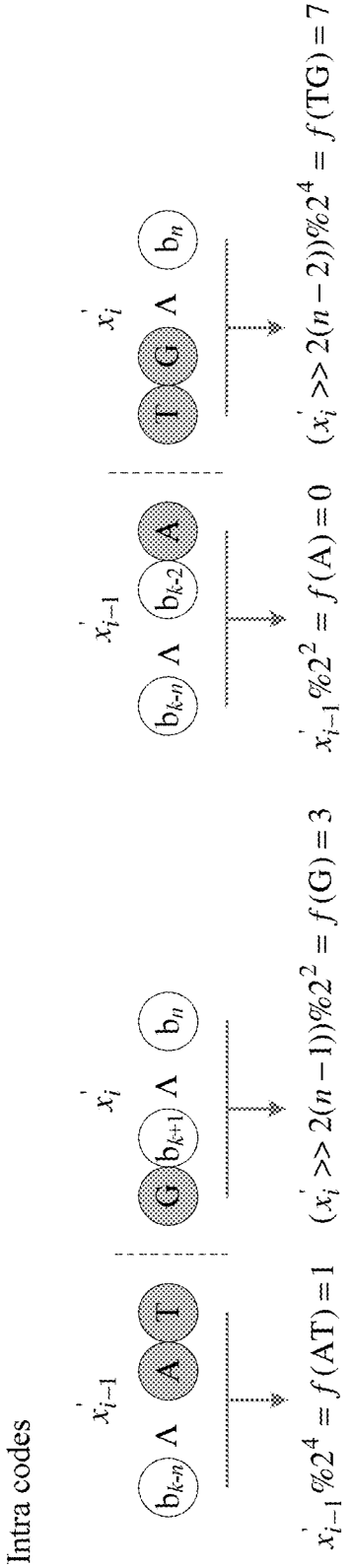

Intra codes
x'_{i-1}
x'_i
b_{k-n} Λ A T G b_{k+1} Λ b_n
$x'_{i-1}\%2^4 = f(\mathrm{AT}) = 1$
$(x'_i >> 2(n-1))\%2^2 = f(\mathrm{G}) = 3$
x'_{i-1}
x'_i
b_{k-n} Λ b_{k-2} A T G Λ b_n
$x'_{i-1}\%2^2 = f(\mathrm{A}) = 0$
$(x'_i >> 2(n-2))\%2^4 = f(\mathrm{TG}) = 7$

FIG 3A

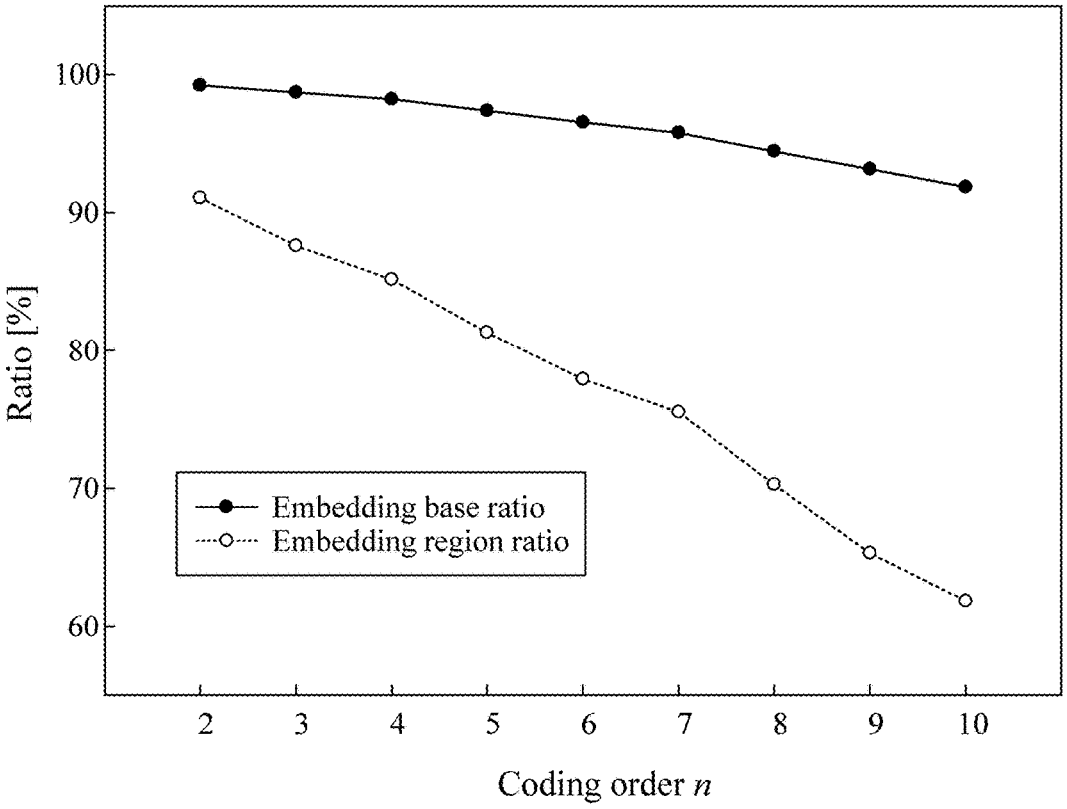

100
90
80
70
60
Ratio [%]
Embedding base ratio
Embedding region ratio
2
3
4
5
6
7
8
9
10
Coding order n

FIG. 3B

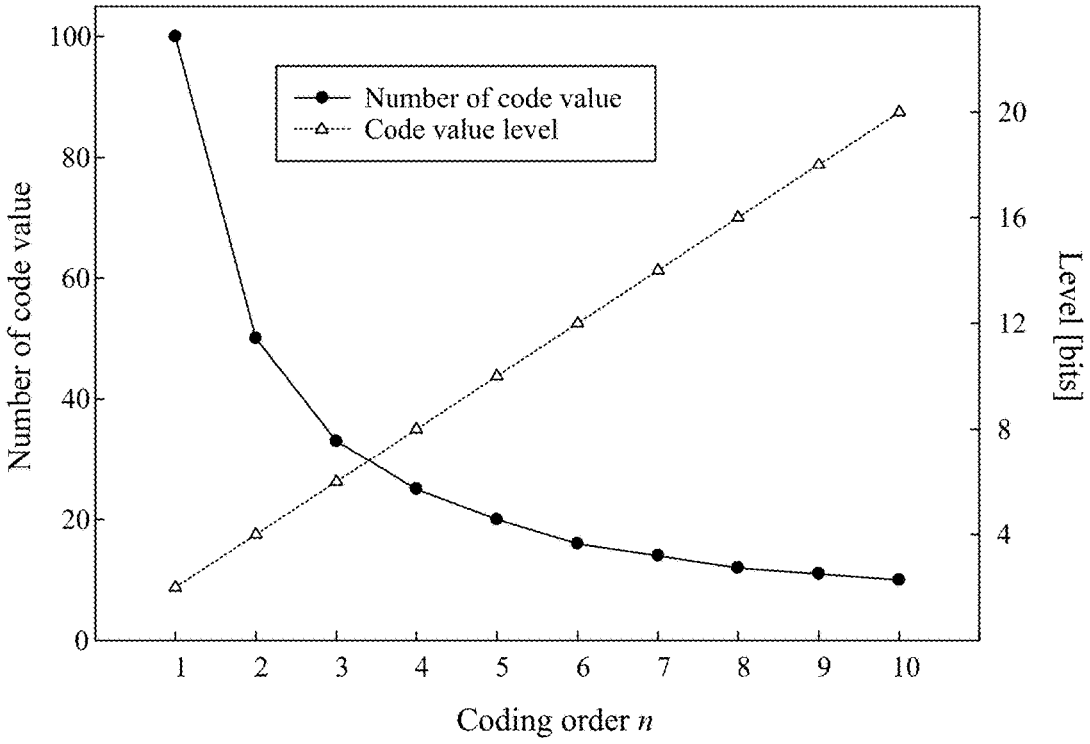

Number of code value
Code value level
Number of code value
Level [bits]
Coding order n
0
20
40
60
80
100
4
8
12
16
20
1
2
3
4
5
6
7
8
9
10

FIG. 4A

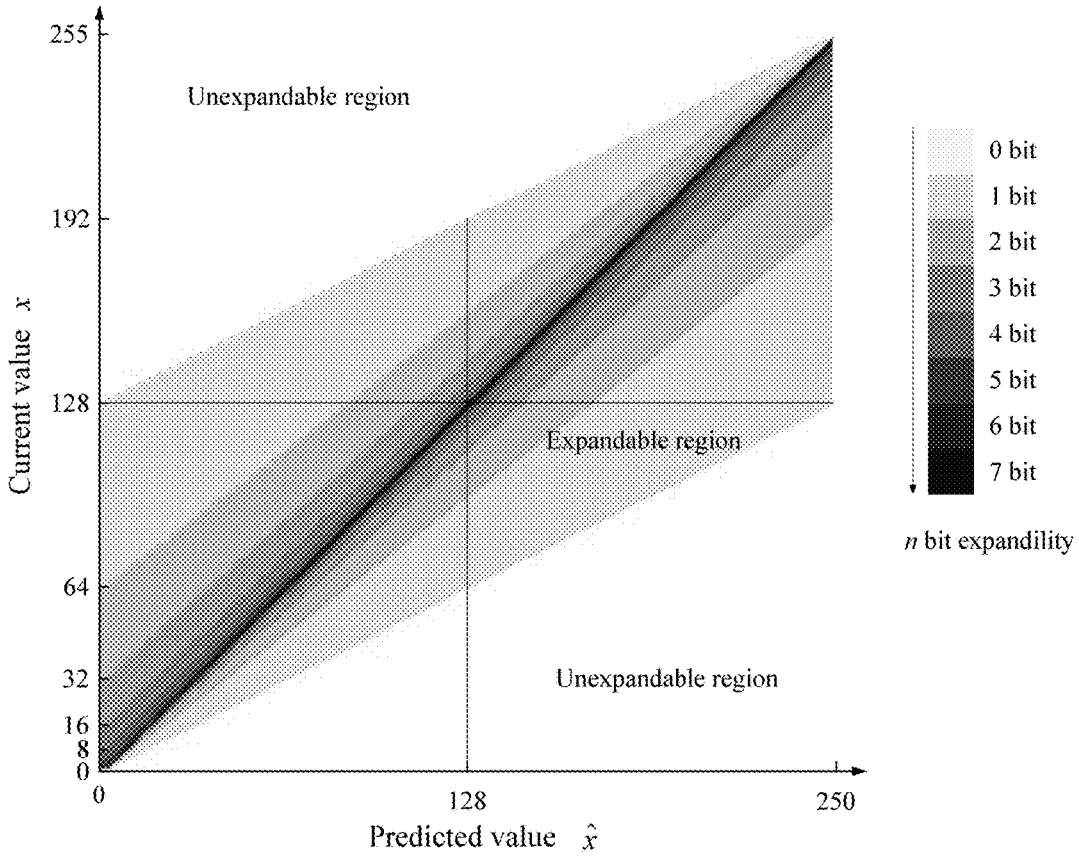

255
192
128
64
32
16
8
0
Current value x
Unexpandable region
Expandable region
Unexpandable region
0
128
250
Predicted value $\hat{x}$
0 bit
1 bit
2 bit
3 bit
4 bit
5 bit
6 bit
7 bit
n bit expandility

FIG. 4B

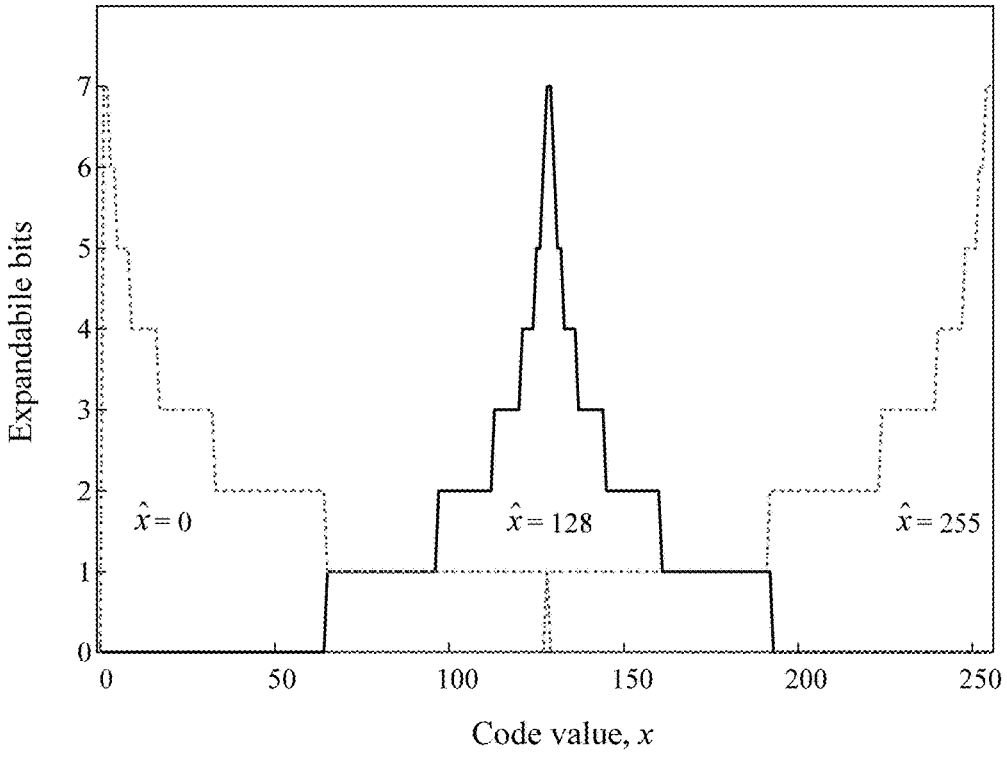
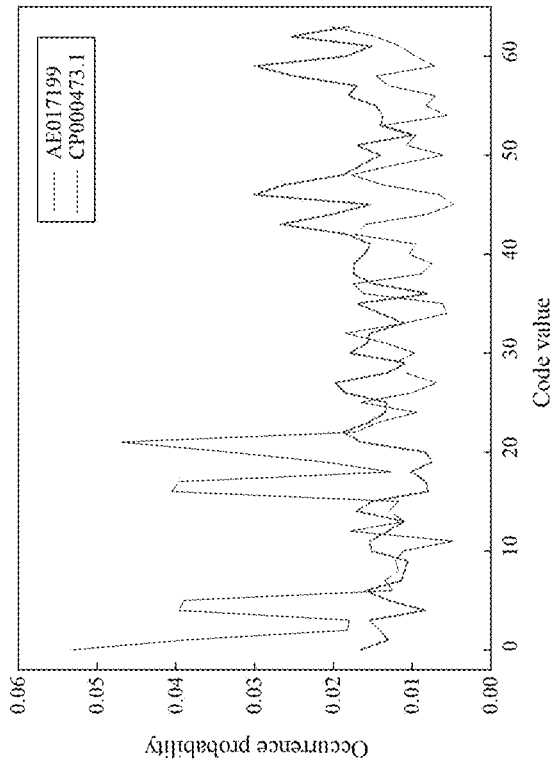

Expandabile bits
0
1
2
3
4
5
6
7
$\hat{x}=0$
$\hat{x}=128$
$\hat{x}=255$
0
50
100
150
200
250
Code value, x

FIG. 5A


AE017199
CP000473.1
Occurrence probability
Code value
0.06
0.05
0.04
0.03
0.02
0.01
0.00
0
10
20
30
40
50
60

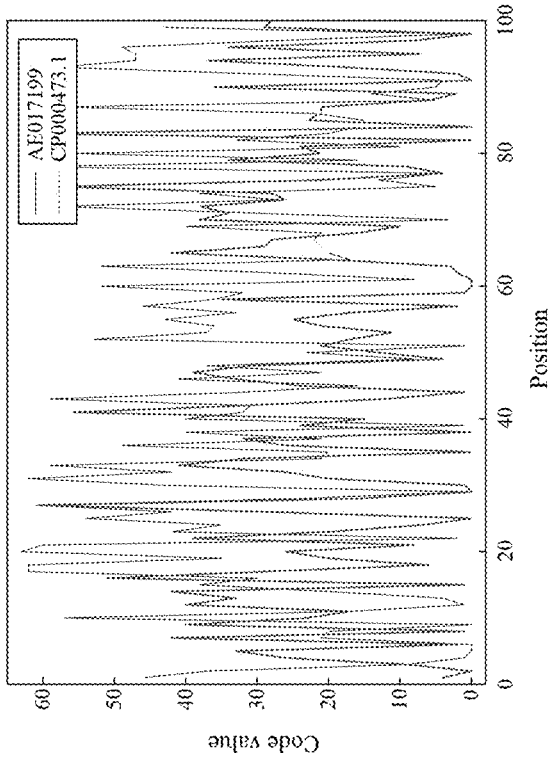
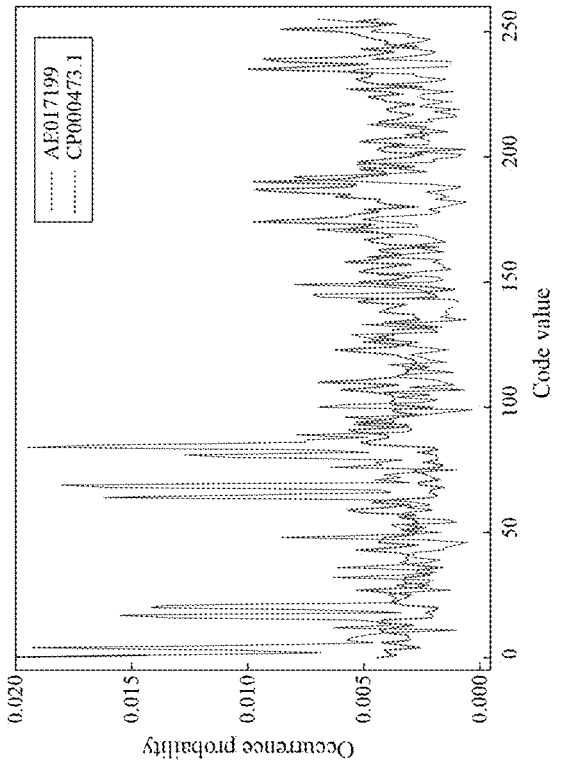

AE017199
CP000473.1
Code value
Position
60
50
40
30
20
10
0
0
20
40
60
80
100

FIG. 5B


AE017199
CP000473.1
Occurrence probability
0.020
0.015
0.010
0.005
0.000
0
50
100
150
200
250
Code value

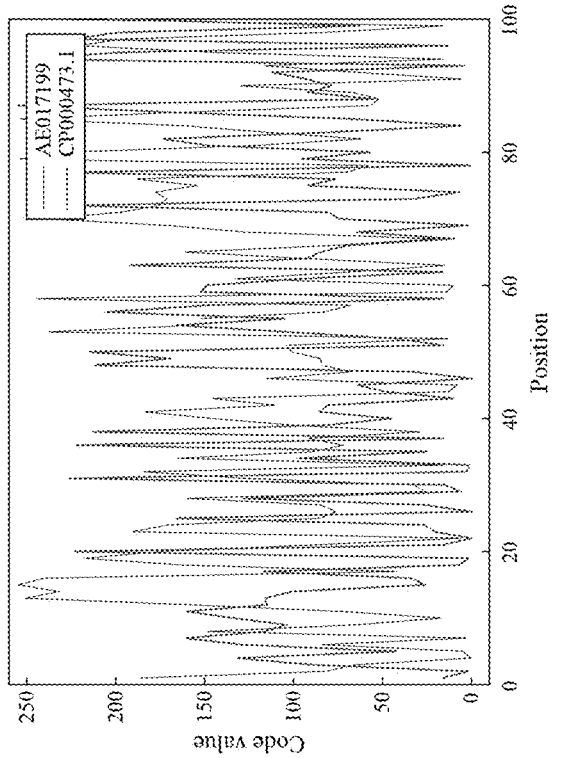

AE017199
CP000473.1
Code value
250
200
150
100
50
0
0
20
40
60
80
100
Position

FIG. 6A

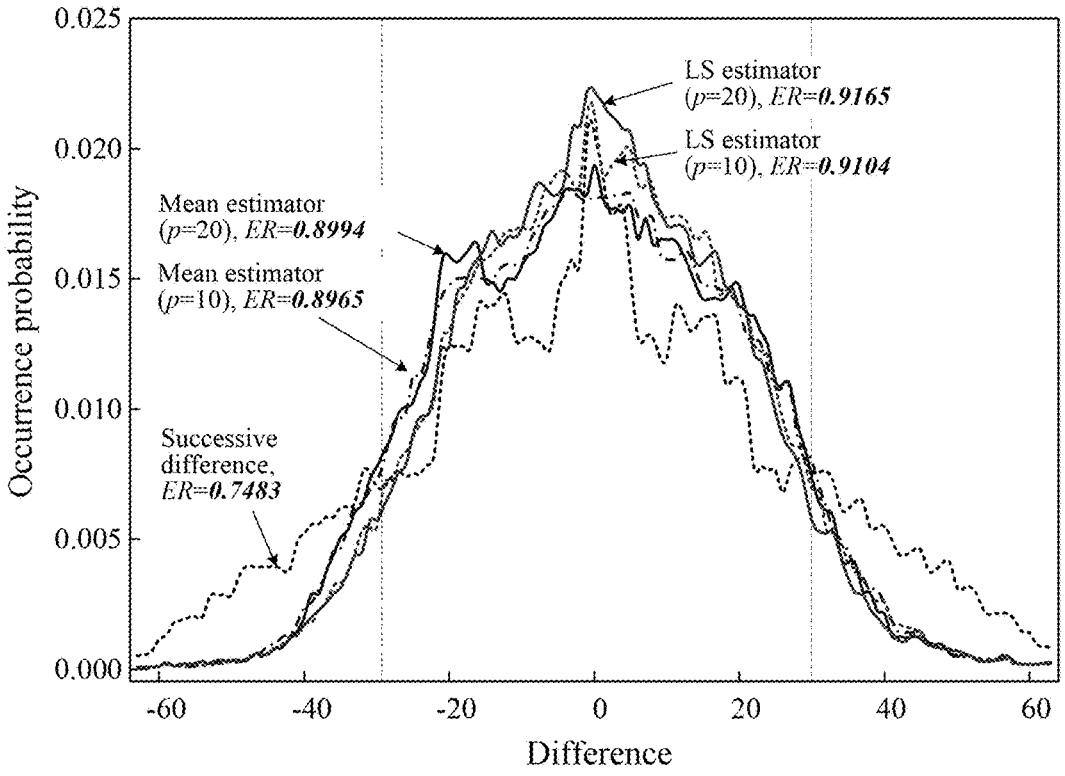

0.025
0.020
0.015
0.010
0.005
0.000
Occurrence probability
-60
-40
-20
0
20
40
60
Difference
LS estimator
(p=20), ER=0.9165
LS estimator
(p=10), ER=0.9104
Mean estimator
(p=20), ER=0.8994
Mean estimator
(p=10), ER=0.8965
Successive
difference,
ER=0.7483

FIG. 6B

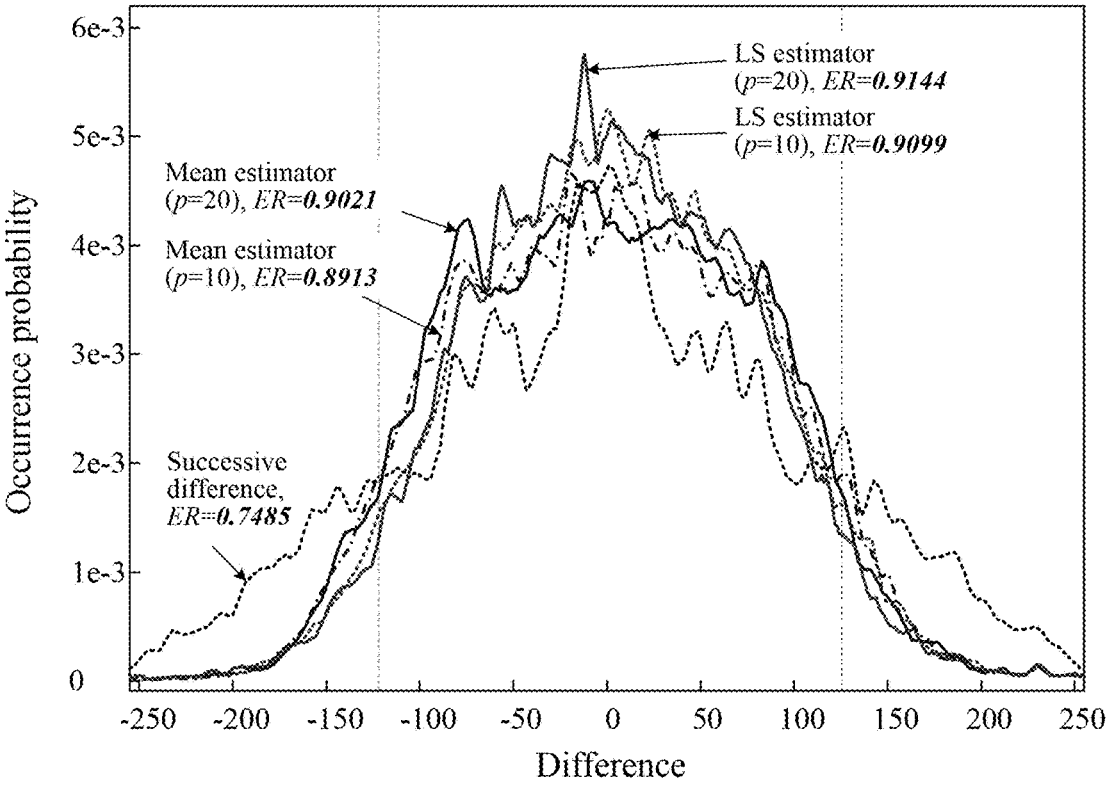

6e-3
5e-3
4e-3
3e-3
2e-3
1e-3
0
Occurrence probability
-250
-200
-150
-100
-50
0
50
100
150
200
250
Difference
LS estimator
(p=20), ER=0.9144
LS estimator
(p=10), ER=0.9099
Mean estimator
(p=20), ER=0.9021
Mean estimator
(p=10), ER=0.8913
Successive
difference,
ER=0.7485

FIG. 7

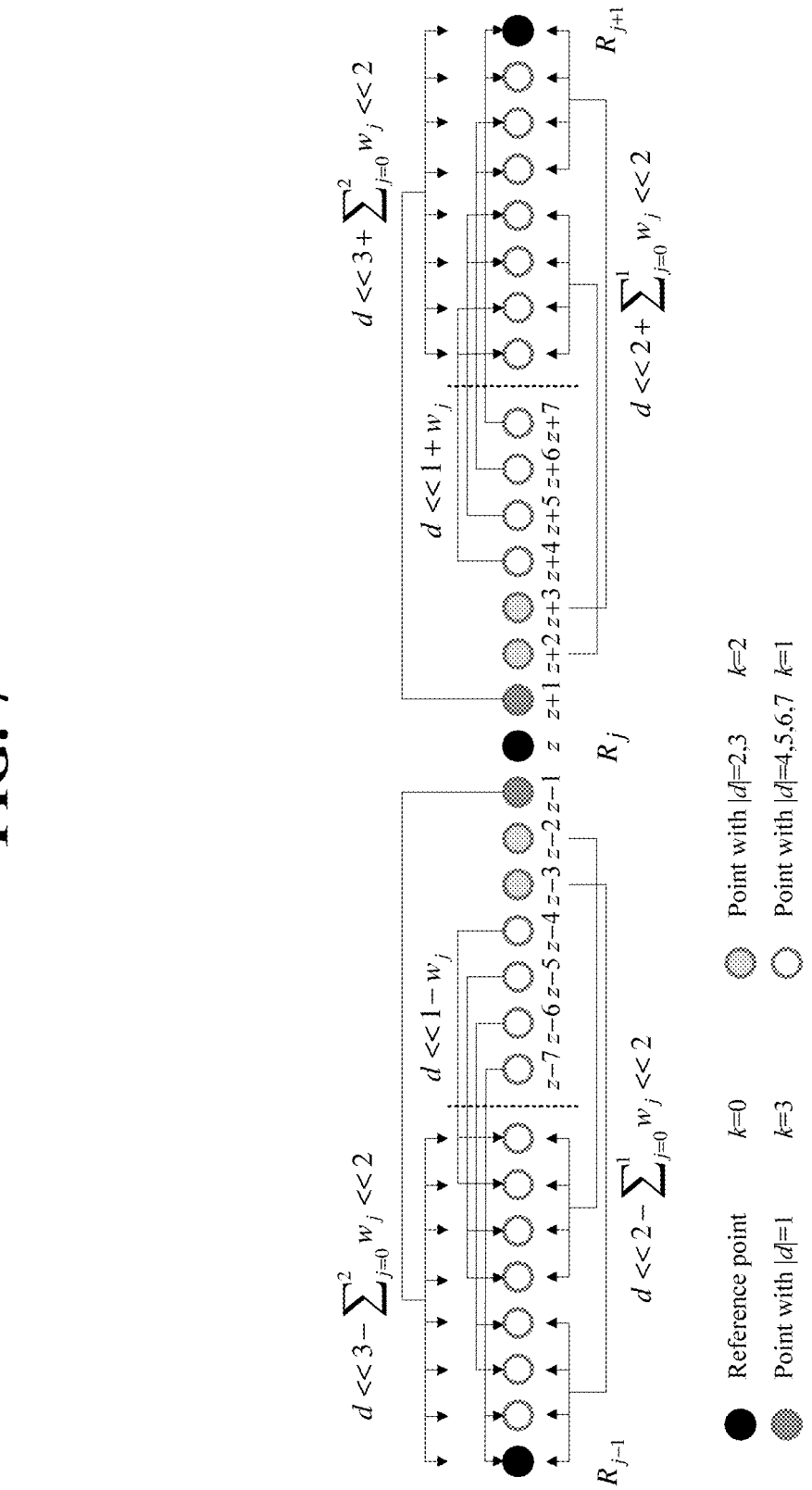

$d << 3 + \sum_{j=0}^{2} w_j << 2$
$d << 1 + w_j$
$d << 2 + \sum_{j=0}^{1} w_j << 2$
R_{j+1}
$z+7\ z+6\ z+5\ z+4\ z+3\ z+2\ z+1\ z\ z-1\ z-2\ z-3\ z-4\ z-5\ z-6\ z-7$
R_j
$d << 3 - \sum_{j=0}^{2} w_j << 2$
$d << 1 - w_j$
$d << 2 - \sum_{j=0}^{1} w_j << 2$
R_{j-1}
Reference point
Point with |d|=1
k=0
k=3
Point with |d|=2,3
Point with |d|=4,5,6,7
k=2
k=1

FIG. 8A

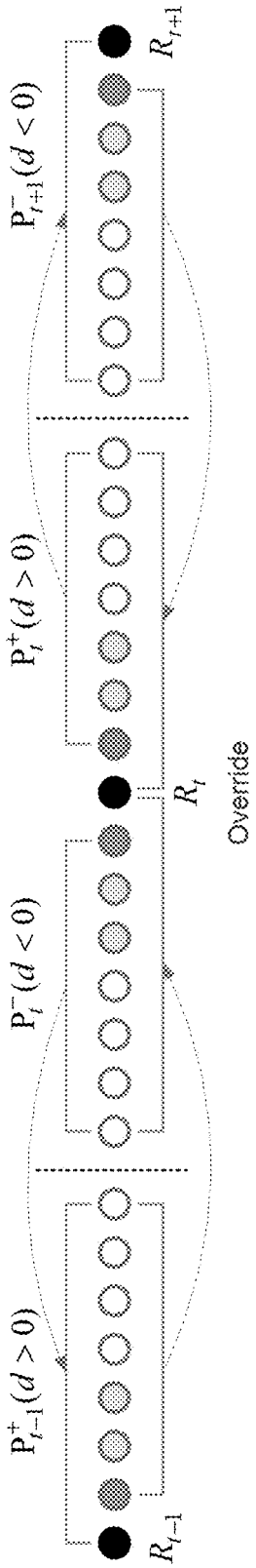


FIG. 8B

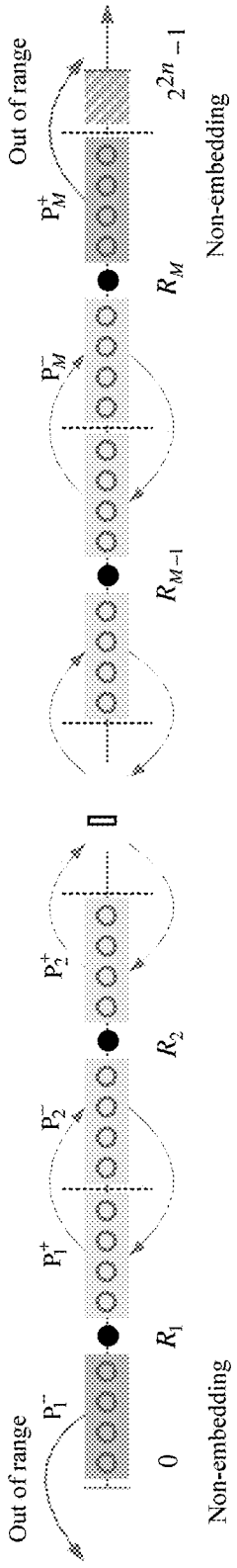
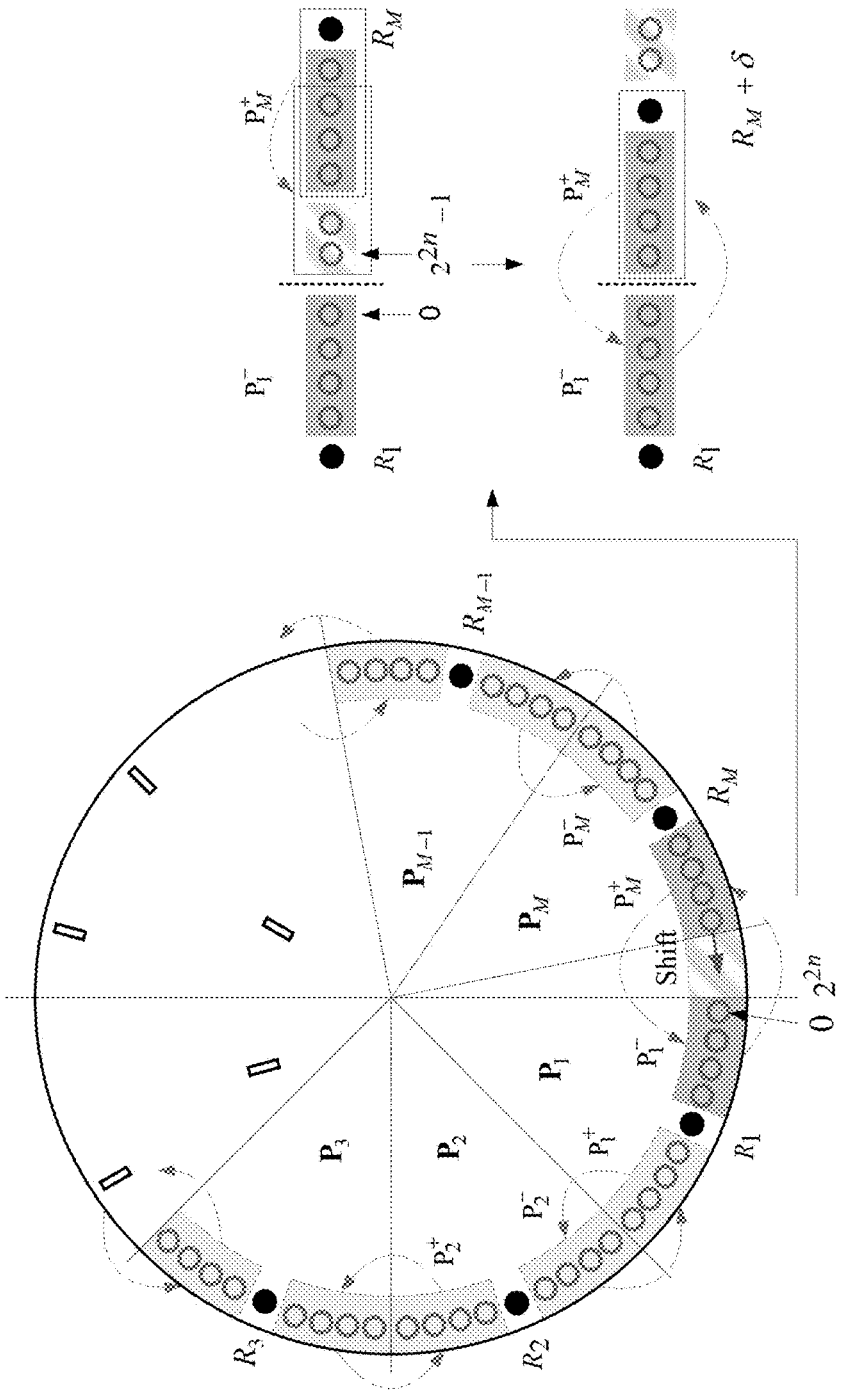

Out of range
P_1^-
0
R_1
P_1^+
P_2^-
R_2
P_2^+
P_{M-1}^-
R_{M-1}
P_{M-1}^+
P_M^-
R_M
P_M^+
$2^{2n}-1$
Out of range
Non-embedding
Non-embedding

FIG. 9


R_M
P_M^+
$2^{2n}-1$
0
P_1^-
R_1
$R_M+\delta$
R_{M-1}
R_3
R_2
P_{M-1}
P_M
P_M^-
Shift
P_1^-
P_3
P_2
P_1
P_1^+
P_2^-
P_2^+
$0\ 2^{2n}$

REVERSIBLE DNA INFORMATION HIDING METHOD BASED ON PREDICTION-ERROR EXPANSION AND HISTROGRAM SHIFTING

CROSS REFERENCE TO RELATED APPLICATION

[0001] The present application claims priority to Korean Patent Application No. 10-2018-017337, filed Feb. 13, 2018, which is incorporated herein by reference.

TECHNICAL FIELD

[0002] The present invention relates generally to a reversible DNA information hiding method based on prediction-error expansion and histogram shifting, the method being capable of false start codon prevention, original sequence length preservation, high watermark capacity, and blind detection based on prediction-error expansion and histogram shifting without biological mutation.

RELATED ART

[0003] A DNA sequence consists of a coding DNA and a non-coding DNA, and watermarks are inserted into the two regions, respectively, such that data can be hidden. In the case of the coding DNA, a redundancy codon range is extremely small, and thus the coding DNA is not suitable for reversible watermarking. In the case of the non-coding DNA, a watermark available range is wide compared to the coding DNA due to no condition for protein code preservation, and thus the non-coding DNA is suitable for DNA reversible watermarking.

[0004] Lossless compression and difference expansion (DE)-based methods widely used in conventional reversible image watermarking have been proposed by T. Chen, et al. (reference [1]). A histogram-based reversible DNA watermarking method with a low modification rate of bases has been proposed by Huang, et al. (reference [2]). In this method, the modification rate of bases is low, but bpn is extremely low and a false start codon occurs, similar as Chen's method.

[0005] Furthermore, a piecewise linear chaotic map (PWLCM)-based information hiding method has been proposed by Liu, et al. (reference [3]). Information hiding methods for tamper location detection and restoration of a DNA sequence have been proposed by J. Fu (reference [4]) and Ma (reference [5]). These methods are for hiding data using substitution by complementary rule, and non-blind methods requiring a reference (or original) DNA sequence for extraction and restoration.

[0006] The foregoing is intended merely to aid in the understanding of the background of the present invention, and is not intended to mean that the present invention falls within the purview of the related art that is already known to those skilled in the art.

SUMMARY

[0007] Accordingly, the present invention has been made keeping in mind the above problems occurring in the related art, and the present invention is intended to propose a reversible DNA information hiding method based on prediction-error expansion and histogram shifting, the method being capable of false start codon prevention, original sequence length preservation, high watermark capacity, and blind detection based on prediction-error expansion and histogram shifting without biological mutation.

[0008] In order to achieve the above object, according to one aspect of the present invention, there is provided a reversible DNA information hiding method based on prediction-error expansion and histogram shifting, the method including: coding, at a first step, a four-letter base sequence of a non-coding region DNA to an n order code value; embedding, at a second step, multiple bits for each code value by a least square (LS) prediction error; embedding, at a third step, an n order watermark bit by non-circular histogram and circular histogram multi-level shifting; verifying, at a fourth step, occurrence of a start code of a watermarked intra code value and a watermarked inter code value.

[0009] At the first step, b may be a four-letter base b={'A', 'T', 'C', 'G'}, b may be a base value of the b, x may be a base block consisting of n bases, x may be a code value for the base block x, and n may be a coding order. Coding to a 2n-bit code value x in units of the base block x consisting of the n bases may be performed as follows

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

where $x=(b_1, b_2, \ldots, b_n)$, $x \in [0, 2^{2n}-1]$. The bases of the base block may be restored from the code value x as follows $f^{-1}(x)=x$ where $b_k=(x>>2(n-k))\%4$ for $k=1, \ldots, n$.

[0010] At the fourth step, preventing of a false start codon in the watermarked intra code value may include: generating a code value table containing the false start codon in advance; and embedding a watermarked code value not to contained in the code value table.

[0011] At the fourth step, preventing of a false start codon in the watermarked intra code value may include: when a previous watermarked code value x'_{i-1} is given, a number of embedded bits for a current processed code value is controlled such that the current processed code value x'_i does not satisfy

$$x'_{i-1}(n-1,n) \| x'_i(1,2) \in Z^c$$

[0012] if $(x'_{i-1}\%2^4)=f('AT')=1$ and $(x'_i>>2(n-1))\%2^2=f('G')=3$

[0013] if $(x'_{i-1}\%2^2)=f('A')=0$ and $(x'_i>>2(n-2))\%2^4=f('YG')=7$.

[0014] At the second step, the code value may be predicted through local prediction for each embedding region.

[0015] The present invention has been made keeping in mind the above problems occurring in the related art. According to the reversible DNA information hiding method based on prediction-error expansion and histogram shifting, false start codon prevention, original sequence length preservation, high watermark capacity, and blind detection based on prediction-error expansion and histogram shifting are possible without biological mutation

BRIEF DESCRIPTION OF THE DRAWINGS

[0016] The above and other objects, features and other advantages of the present invention will be more clearly understood from the following detailed description when taken in conjunction with the accompanying drawings, in which:

[0017] FIGS. **1**A and **1**B are views illustrating a general 2-bit base value and a 2n-bit value for n order base blocks, respectively;

[0018] FIGS. **2**A and **2**B are views illustrating occurrence probability of a false start codon in an intra code value and in inter code values, respectively;

[0019] FIGS. **3**A and **3**B are views illustrating, with respect to the coding order n with x=1, a ratio $R_{region}(n)$ of the number of embedding regions and a ratio $R_{base}(n)$ of the number of bases, and a code value level and the number of code values when the number of bases is 100;

[0020] FIGS. **4**A and **4**B are views illustrating an expandable region of x for a prediction value $\hat{x}$, and the number of expandable bits of x with the prediction value $\hat{x}$=0, 128, 255

$$\alpha(k) = \mathrm{sgn(d)}\sum_{i=0}^{k-1} 2^j \omega_{j+1},$$

when all watermark bits have values of one, $w=\{1\}_1^{2n-1}$.

[0021] FIGS. **5**A and **5**B are views illustrating code values of 'AE017199' and 'CP000473.1' sequences, histograms of the code values, successive predictor difference histograms when the coding orders are n=3 and n=4;

[0022] FIGS. **6**A and **6**B are views illustrating mean error histograms of LS predictors, mean predictors, and successive predictors of 'AE017199', 'CP000473.1' sequences when the coding orders are n=1 and n=4;

[0023] FIG. **7** is a view illustrating shift of values where differences from a center value R_i are d>0 and d<0 on an arbitrary section P_i of an n order code value histogram domain Z;

[0024] FIGS. **8**A and **8**B are views illustrating code value shifting on a current section P_i and left and right adjacent sections P_{i-1} and P_{i+1}, and code value shifting between each section and left and right adjacent sections on the entire sections; and

[0025] FIG. **9** is a view illustrating data hiding based on circular histogram shifting.

DETAILED DESCRIPTION

[0026] According to a preferred embodiment of the present invention, a reversible DNA information hiding method based on prediction-error expansion and histogram shifting is a method using difference expansion (DE) of a multi-bit base code value and histogram shifting, and main features of the present invention are as follows.

[0027] 1. Blind Reversibility: a reversible watermark is hidden without change in the length of a DNA sequence and in amino acid, and extraction and restoration are possible without an original DNA sequence.

[0028] 2. Watermarking Usability: a base bit sequence of a bit is encoded to a code value sequence of 2n bits, such that reversible watermark hiding, extraction, and restoration processes are easily performed.

[0029] 3. Watermark Capacity: based on DE and histogram shifting of a code value sequence, multi-bit embedding for each target code value is enabled, and thus watermark capacity is increased.

[0030] 4. No false start codon: through a false start codon—code value table and comparison-search between adjacent code values, occurrence of a false start codon in an intra code value and inter code values is prevented.

[0031] Before description of the present invention, symbols used in the present invention are defined as follows.

[0032] A DNA sequence consists of a non-coding region D^{nc} and a coding region D^c.

[0033] The non-coding region D^x is divided into an embedding region Γ and a non-embedding region $\Gamma^C=D^{nc}-\Gamma$.

[0034] An embedding target region Γ has regions D_i of |Γ| numbers, and each region D_i consists of bases of $|D_i|$ numbers; $\Gamma=\{D_i\}_{i=1}^{|\Gamma|}$, $D_i=\{b_j\}_{i=1}^{|D_i|}$.

[0035] b is a four-letter symbol base b={'A', 'T', 'C', 'G'}, and b is a base value of b.

[0036] $x=\{b_1, b_2, \ldots, b_n\}$ is a base block consisting of n bases, and x is a code value for the base block x. Here, n is called a coding order.

[0037] x' is a watermarked code value, and $x'=\{b'_1, b'_2, \ldots, b'_n\}$ is a base block of x'.

[0038] $W=\{w_1, w_2, \ldots, w_{N_w}\}$, w∈[0,1] is a watermark bit string to be hidden.

[0039] Cardinality |D| of a matrix L indicates the number of elements or length of L.

[0040] 1. Coding of Four-Letter Base

[0041] For ease of watermarking signal processing on a four-letter base sequence, multi-bit coding processing is essential. In this section, the multi-bit coding processing for ease of watermarking signal processing and false start codon prevention will be described.

[0042] 1-1. Coding Based on a Coding Order

[0043] Generally, a nucleotide base is expressed as four letters, b=(A, T, C, G) as shown in FIG. **1**A, that are expressed as four decimal numbers or 2-bit binary numbers.

$$b=(0,1,2,3)_{10}=(00,01,10,11)_2 \leftarrow b=(A,T,C,G) \quad (1)$$

[0044] For ease of signal processing, rather than a 2-bit value, as shown in FIG. **3**B, expansion to a value expressed in multiple bits of two or more bits is required. In the present invention, coding to a 2n-bit code value x in units of a base block x consisting of n bases is performed as follows.

$$x = f(x) = \sum_{k=1}^{n} (b_k \cdot 2^{2(n-k)}) \text{ where}$$
$$x = (b_1, b_2, \ldots, b_n), x \in [0, 2^{2n} - 1] \quad (2)$$

[0045] The bases of the base block are easily restored from the code value x as follows.

$$f^{-1}(x)=x \text{ where } b_k=(x>>2(n-k))\%4 \text{ for } k=1, \ldots, n \quad (3)$$

[0046] In the present invention, the number n of bases of the base block is called a coding order. Bases in the embedding region D_i are coded to a code value X_i based on the coding order n; $X_i=\{x_k|k\in[1,N_i]\}$, $N_i=\lfloor|D_i|n\rfloor$. Here, the number N_i of code values is determined by the coding order n.

[0047] 1-2. False Start Codon Prevention

[0048] The false start codon may occur in an intra code value or inter code values as follows.

[0049] 1) Intra Code Value

[0050] a code value domain based on the coding order n is $z\in Z=[0,2^{2n}-1]$. In the case of n>2, as shown in FIG. **2**A, false start codons of n−2(n>2) numbers may occur in the code value domain. The number of code values containing false start codons occurring at arbitrary positions j∈[1,n−2]

in the base block is $2^{2(n-3)}$ and thus the total number of code values containing false start codons occurring at n−2 positions is $(n-2)\times 2^{2(n-3)}$. The code value containing the false start codon z' is defined as follows.

$$z^C = \sum_{k=1}^{j-1} b_k 2^{2(n-k)} + 0\times 2^{2(n-i)} + 1\times 2^{2(n-j+1)} + 3\times 2^{2(n-j+2)} + \sum_{k=j+3}^{n} b_k 2^{2(n-k)} \quad (4)$$

[0051] for ∀j=[1,n−2] and $\forall b_k \in [A,T,C,G]$, k=1, 2, . . . , j−1, j+3, . . . , n

[0052] Here, the symbols 'A', 'T', and 'G' correspond to 0, 1, and 3 as shown in Formula (3), and except for consecutive bases {A,T,G} on arbitrary positions, all bases at remaining positions have {A, T, C, G}. According to the present invention, in coding of the base, a code value table $Z^c=\{z^c\}$ including the false start codon is generated in advance, and then an embedding process is performed for a watermarked code value x' not to be included in the Z.

[0053] 2) Inter Code Values

[0054] The false start codon may occur between a base block x'_{i-1} of a previous watermarked code value x'_{i-1} and a base block x'_1 of a current processed code value x'_1. As shown in FIG. **2**B, in the case of $(x'_{i-1}\ x'_i)$, when (. . . A, TG . . .) or (. . . AT, G . . .) the false start codon occurs in the middle portion thereof. Thus, two code values including the false start codon therebetween are defined as follows.

$$x'_{i-1}(n-1,n)\|x'_i(1,2)\in Z^c \quad (5)$$

[0055] if $(x'_{i-1}\%2^4)=f(\text{'AT'})=1$ and $(x'_i>>2(n-1))\%2^2=f(\text{'G'})=3$

[0056] if $(x'_{i-1}\%2^2)=f(\text{'A'})=0$ and $(x'_i>>2(n-2))\%2^4=f(\text{'YG'})=7$.

[0057] x(j,j+1) indicates the j-th and j+1-th bases of the code value x, and || indicates a concatenation operator. $x'_{i-1}(n-1,n)\|x'_i(1,2)$ indicates a code value where the n−1-th and n-th bases of x'_{i-1} are concatenated with the first and second bases of x'_i. In the present invention, when the previous watermarked code value x'_{i-1} is provided, the number of embedded bits for the code value x_i is controlled to prevent the current watermarked code x'_i from satisfying the above condition.

[0058] 2. Embedding Region (Target Region) Selection

[0059] In the present invention, a watermark is embedded into a code value string generated in units of a base block. Here, a region with a short sequence length is not suitable for a watermark embedding target due to a short code value string. Thus, the embedding region is a region having a or more code values, and a set Γ(n) of embedding regions for the coding order n is defined as follows.

$$\Gamma(n)=\{D_i \mid |D_i|>\alpha p\times n\}, D_i=\{b_{ii} \mid j\in[1,|D_i|]\} \quad (6)$$

[0060] Here, D_i indicates the i-th embedding region, b_{ii} indicates the j-th four-letter base in the D_i region, and $|D_i|$ indicates the number of bases in D_i. α indicates the minimum number of code values in the embedding region, and x indicates a prediction order, which will be described in section 3. According to an embodiment of the present invention, the minimum value of code values is set to 10 or more, and the embedding region is selected based on the prediction order x.

[0061] A ratio of the number of embedding regions to the total number of non-coding regions on the given DNA sequence is designated by $R_{region}(n)$, and a ratio of the number of bases in embedding regions to the number of bases in total non-coding regions is designated by $R_{base}(n)$. FIG. **3**A shows the ratio $R_{region}(n)$ of the number of embedding regions and the ratio $R_{base}(n)$ of the number of bases when the coding order n ranges 2 to 10 on the DNA sequence. FIG. **3**B shows the code value level with respect to the coding order n and the number of code values, when the number of bases is 100. Referring to these figures, $R_{region}(n)$ decreases in proportion to increase of n, but $R_{base}(n)$ is maintained at 92% or more. In the case where the number of bases is given, when n increases, the number of code values geometrically decreases, but the code value level increases. That is, when the code value level is high, the range of watermarking signal processing is wide and the number of bases is maintained, but the number of target code values is small, and thus watermark capacity is limited. In the present invention, since multiple bits per code value are embedded, when the code value level increases, the number of embedded bits per code value increases, but the number of code values decreases. Thus, on the given non-coding region, the optimum coding order n for the watermark capacity is required.

[0062] 3. Code Value Prediction-Error Expansion (PE)-Based Reversible Watermarking

[0063] When a code value of the non-coding region is given, a prediction-error expansion method used in a conventional image data may be used to embed a bit in a pair of code values. For example, when a prediction x̂ value a with respect to an arbitrary code value x and a watermark bit w are given, the embedded code value x' is as follows.

$$x'=\ddot{x}+2(x-\ddot{x})+w=2x-\ddot{x}+w \quad (7)$$

[0064] Watermark extraction and code value restoration are easily obtained from ẍ and x' as

$$w = x' - \hat{x} - 2\left\lfloor \frac{x'-\hat{x}}{2} \right\rfloor, \quad x = \frac{1}{2}(x' + \hat{x} - w).$$

This method is suitable for image data with high correlation between adjacent pixels. By a prediction error modeled as Laplacian distribution, one bit can be embedded into each of pixel pairs.

[0065] However, code values of the DNA sequence have a low correlation between successive predictors, and thus an adaptive prediction is required. Also, code values can be moved without limitation under false start codon limitation conditions, and thus multiple bits can be embedded in a pair of code values. Thus, in this section, a code value prediction-error expansion-based multi-bit embedding method will be described.

[0066] 3-1. Code Value Error Expansion Condition for Multi-Bit Embedding

[0067] Except for false start codon values, DNA code values having no condition for definition move without limitation within a valid range. Thus, the prediction error d for a pair of code values can be expanded 2^k times according

to an expansion condition to embed k bits, and at most 2n−1 bits can be embedded; k_{max}=2n−1.

[0068] When k bits of watermark $\{w_j\}_1^k$ and a prediction value x̂ are given, a k-bit embedded code value x' is obtained by the 2^k times expanded prediction error d as follows.

$$x' = \hat{x} + 2^k d + \mathrm{sgn(d)} \sum_{i=1}^{k} 2^{j-1} w_1 \text{ where } d = x - \hat{x} \tag{8}$$

[0069] When the embedded code value x' and the number k of bits are given, watermark extraction and restoration are easily performed as follows.

$$w_i = ((x' - \hat{x}) >> (j-1)) \% 2 \text{ for } j = 1, \ldots, k \tag{9}$$

$$x = \hat{x} + d = \hat{x} + (x' - \hat{c}) >> k \tag{10}$$

[0070] Since the embedded code value x' is desired to be $0 \le x' \le 2^{2n}-1$, expansion condition of the prediction error d for 2^k times expansion is as follows.

$$2^{-k}\left(-\hat{x} - \mathrm{sgn}(d)\sum_{i=1}^{k} 2^{j-1} w_j\right) \leqq d \leqq 2^{-k}\left(2^{2n} - 1 - \hat{x} - \mathrm{sgn}(d)\sum_{i=1}^{k} 2^{j-1} w_j\right) \tag{11}$$

[0071] The code value x is desired to satisfy the condition as follows.

$$x \in [\max(0, \lceil \hat{c} + 2^{-k}(-\hat{x} - \alpha(k)) \rceil), \min(2^{2n}-1, \lfloor \hat{x} + 2^{-k}(2^{2n} - 1 - \hat{x} - \alpha(k)) \rfloor)], \tag{12}$$

[0072] where

$$\alpha(k) = \mathrm{sgn(d)} \sum_{i=1}^{k} 2^{j-1} w_j.$$

Such the expansion condition is determined depending on watermark k bits and $\{w_j\}_1^k$ the prediction value x̂, and the number of bits to be embedded in the code value x is determined depending on the expansion condition.

[0073] FIG. **5**A shows the number of bits to be embedded in the code value x for each prediction value x̂ when the coding order is n=4 ($x, \hat{x} \in \lceil 1, 2^s - 1 \rceil$) and all watermark bits are 1 w={1}. The maximum number k_{max} of embedded bits is 2n−1=7. FIG. **5**B shows a range of code values x depending on the number of embedded bits when the prediction value x̂ is 0, 128, and 255. When the number of embedded bits is large, an expandable region is geometrically narrow, and when x̂ is close to 0 or 255, the number of embedded bits is small.

[0074] 3.2 Code Value Prediction

[0075] FIGS. **5**A and **5**B show code values and code value histograms of 'AE017199' and 'CP000473.1' sequences, when the coding orders n are 3 and 4. The code value histogram is expanded or reduced depending on the coding order, but distribution is not standardized depending on the sequence. That is, code values of the 'AE017199' sequence are evenly distributed in, except for four regions, the remaining regions, and code values of the 'CP000473.1' sequence are evenly distributed with white noise in the whole regions. Also, the code value sequence appears in random form, and correlation between successive predictors is extremely low. Thus, in the present invention, in order to reduce the prediction error for the code value, the code value is predicted based on a local LS predictor, such as Dragoi, etc.

[0076] A row vector of x code values for predicting the current code value x_i is $x_i=(x_{i-1}, \ldots, x_{i-v})$ and a row vector of x parameter is $b=(\beta_1, \ldots, \beta_v)$. Here, x indicates a prediction order. When x_i is observed, the prediction value $\hat{x}_1$ of x_1 is defined by a linear regression function $f_\beta(x)$ as follows.

$$\hat{x}_i = f_\beta(x_i) = \sum_{i=1}^{p} \beta_j x_{i-j} = x_i b' \tag{13}$$

[0077] When a row vector of all code values in an arbitrary embedding region is $y=(x_1, \ldots, x_N)$ and N×p matrix of N observed previous code values is $X=(x'_1, \ldots, x'_N)$, LS predictor computes parameter t that minimizes the square distance) $\|y'-Xb'\|^2=(u'-Xb')'(u'-Xb')$ between u' and Xb' as follows.

$$b=(X'X)^{-1}X'y' \tag{14}$$

[0078] In the present invention, rather than whole prediction on whole embedding regions, local prediction for each embedding region is performed to predict the code value. Thus, in decoding process, additional information of |Γ(n)|×t which is parameter t by the number |Γ(n)| of embedding regions of the DNA sequence is required.

[0079] The code value may be predicted using a successive predictor $\hat{x}_i=x_{i-1}$ or a mean predictor

$$\hat{x}_i = \sum_{i=1}^{p} x_{i-j} / p.$$

FIGS. **6**A and **6**B show prediction error histograms for successive predictors, mean predictors, and LS predictors when the coding orders are n=3 and n=4 for 'AE017199' and 'CP000473.1' sequences (p is a prediction order (the number of successive predictors used in prediction), and ER (expandable region) is expansion region occurrence probability).

[0080] In FIG. **8**, ER indicates expansion region occurrence probability. A successive predictor error has an ER of about 74.8% regardless of the coding order. The mean predictor and the LS predictor have relatively high ER in the case of the coding order n=3, and when the prediction order x is high, ER is high. Particularly, in the case of n=3 and x=20, the LS predictor has the highest ER of 91.6%. That is, in the case of n=3, when the prediction order x of LS is high, insertion capacity is large.

[0081] The prediction error histogram of an image is modeled as Laplacian distribution, but the LS prediction error histogram of the code value is modeled as normal distribution that (μ,σ)=(0,20) with n=3 and x=10, (μ,σ)=(0, 19) with n=3 and x=20, (μ,σ)=(0,80) with n=4 and x=10, and (μ,σ)=(0,76) with n=4 and x=20.

[0082] 3.3 Coding Process

[0083] In the coding process of the present invention, when the coding order n and the prediction order are given, an LS prediction parameter t is obtained for each embedding region. The LS predictor by t is used for the code value x_i with i>p, and the mean predictor is used for the code value with i≤x, thereby obtaining $\hat{x}_1$.

$$\hat{x}_i = \begin{cases} \sum_{j=1}^{p} \beta_j x_{i-j}, & \text{if } i > p \\ \sum_{j=1}^{i-1} \frac{x_{i-j}}{i-1}, & \text{if } 1 < i \leqq p \\ 0, & \text{if } i = 1 \end{cases} \tag{15}$$

[0084] After determining the number k_i ($0 \leq k_i \leq 2n-1$) of embedded bits based on expansion condition of the prediction error $d_i=x_i-\hat{x}_1$, k_1 bits $\{w_I\}_{I=1}^{k_1}$ are embedded in the code value x_1 as follows.

$$x'_i = \hat{x}_i + 2^k_i d_i + \alpha(k_i) \text{ where } \alpha(k_i) = \text{sgn}(d_i) \sum_{l=1}^{k_i} 2^{l-1} w_l \tag{16}$$

[0085] $x'_i \notin Z^t$ and $x'_{i-1}(n-1,n) \| x'_i(1,2) \notin Z^t$

[0086] When the embedded code value x'_1 is included in a false start codon tale Z^t or the previous code value x'_{i-1} includes the false start codon, the number k_i of embedded bits is reduced by one, and then the above-described process is repeated until k_i is zero. In this way, multiple bits are embedded in code values of all embedding regions, and then a watermarked region Γ'(n) is obtained. When k_i is 0, it indicates a non-embedding region of the prediction error or a case where the false start codon occurs.

[0087] The number $K=\{k_i\}$ of embedded bits for each code value and the prediction parameter t for each embedding region are additional information required in watermark extraction and original sequence restoration. It is required that the additional information is included in the watermarked region Γ'(n) and is transmitted without occurrence of the false start codon and generation of another additional information. In the present invention, by arithmetic coding, lossless compression is performed on the number K of embedded bits, the prediction parameter t, and an LSB bit E of a 2-bit base binary number in Γ'(n), thereby generating a compression bit string $C=\{c_i\}$. The compression bit c_i is substituted to the LSB of the binary number b'_i of the four-letter base as follows.

$$b'_i=(b'_i>>1)<<1+c_1, \text{ if } b'_{i-2} \neq 'A' \text{ and } b'_{i-1} \neq 'T' \tag{17}$$

[0088] Here, in a case where two previous embedded bases (b'_{1-2},b'_{1-1}) are "AT", when the current base is b'_1='G', b'_1 is substituted by one of 'A', 'T', and 'C'. When $b'_1 \neq$'G', embedding is omitted. Finally, a base string "AT" in the embedding region Γ"(n) including a compression string C performs as a marker directly indicating that a subsequent base does not include a compression bit. The length of the compression string C is determined by a compression algorithm, but in the present invention, arithmetic coding which is a general lossless compression algorithm is used. Consequently, the DNA sequence $D'=D'^{nc}+D^c$, $D'^{nc}=\Gamma''(n)+\Gamma^c(n)$ containing the additional information and the non-coding region Γ"(n) where the watermark is embedded is transmitted.

[0089] 3.4 Decoding and Restoration Processes

[0090] In decoding process, in the non-coding region Γ"(n) of the DNA sequence D' transmitted first, from the LSB of all bases except for the base following "AT", the number K of embedded bits of the additional information compression string C, the prediction parameter t, and the base LSB bit E are obtained. The code sequence X' of Γ'(n) where the base LSB bit E of Γ"(n) is substituted is obtained by the coding order n. From all code values in X', the watermark is extracted by the number K of embedded bits and the prediction parameter t, and the original code value is restored.

[0091] For example, when the number of embedded bits $k_i>0$ and arbitrary code value x'_i are given, the prediction value $\hat{x}_1$ is obtained from the previous restored code value $(x_{i-1}, \ldots, x_{i-v})$, and then the watermark k_1 bit is extracted from the prediction error $d_i=x'_i-\hat{x}_1$, $w_1=((x'_i-\hat{x}_i)>>(l-1))\%2$ for $l=1, \ldots, k_i$. The original code value x_i is restored by k_i bit shifting of the prediction error d_i as $x_i=\hat{x}_i+((x'_i-\hat{x}_i)>>k_i)$.

[0092] 3.5 Watermark Capacity and Additional Information Amount

[0093] Watermark capacity is affected by the coding order n and the prediction order x. When n and x are given, the number of watermark bits embedded in the embedding region $\Gamma(n)=\{D_i\}_{i=1}^{|\Gamma(n)|}$ is the sum of the number K of embedded bits for each code value in the region. Thus, the number of bits per base (bpn) $bpn_{FE}(n,p)$ is as follows.

$$bpn_{PE}(n, p) = \frac{1}{|\Gamma(n)|} \sum_{i=1}^{|\Gamma(n)|} \left(\frac{1}{N_i} \sum_{i=1}^{N_i} k_j \right) \text{[bit/base]} \tag{18}$$

[0094] where $N_i=\lfloor |D_i|/n \rfloor$ and $0 \leq k_i \leq 2n-1$

[0095] $|\Gamma(n)|$ indicates the number of embedding regions, and N_i indicates the number of code values in the region D_i.

[0096] When $\mathscr{K}$ is LSB substitutable bit amount to embed the additional information compression string C, $\mathscr{K}$ is determined by the number of bases omitted by the false start codon in substituting process. The maximum $\mathscr{K}$ is equal to the total number

$$\sum_{i=1}^{|\Gamma(n)|} |D_i|$$

of bases in Γ'(n). It is required that the length of the additional information compression string C is less than the substitutable bit amount $\mathscr{K}$, the amount of the additional information that is the number K of embedded bits, the prediction parameter t, and the LSB E of 2-bit base is small, or an algorithm with high compression efficiency is required. When an arbitrary watermarked region D'_1 ($\in\Gamma'(n)$) is given, E consists of $|D_i|$ bits, and the number K of embedded bits is expressed by $N_i\lceil \log_2 2n \rceil$ bits, and the prediction parameter t for each embedding region is expressed by x floating points of 32 bits. Thus, additional information $Extra_{PE}(n,p)$ for Γ'(n) is as follows.

$$Extra_{PE}(n,\ p) = \sum_{i=1}^{|\Gamma(n)|} (N_i\lceil \log_2 2n \rceil + |D_i| + 32p)\ [\text{bit}] \tag{19}$$

[0097] When the additional information compression string C is $\rho \times Extra_{PE}(n,p)$, compression is performed to be

$$\rho \times Extra_{PE}(n,\ p) < \Phi \leqq \sum_{i=1}^{|\Gamma(n)|} |D_i|.$$

[0098] 4. Code Value Histogram Shifting-Based Method

[0099] Code values in a non-coding region may be shifted to, except for a code value table having the false start codon, a remaining region. In this section, non-circular and circular code value histogram shifting-based methods for increasing data capacity will be described.

[0100] 4.1 Non-Circular Histogram Shifting (HS)

[0101] (1) Coding Process

[0102] In the present invention, an n order code value histogram domain $Z=\lceil 0,2^{2n}-1 \rceil$ is divided into M sections $\{P_i\}_{i=1}^{M}$. Here, each section is provided in bilateral symmetry with respect to a center value R_i, and R_i is used as a reference value of shifting. Thus, the length of the section has a value of an odd number, and is determined by the number of embedded bits.

[0103] When the maximum number of shifting bits in the section is k_{max} and the center value is $R_i=z$, P_i consists of $2\times 2_{max}^{k}-1$ values as follows.

$$P_i=\{z-2^{k_{max}}+1, \ldots, z-1\cdot z, z+1, \ldots, z+2^{k_{max}}-1\}, \text{for } j\in[1,M] \tag{20}$$

$$R_i=z \tag{21}$$

[0104] The number M of sections is as follows.

$$M = \left\lfloor \frac{2^{2n}}{2\times 2_{max}^{k}-1} \right\rfloor \text{ where } 1 \leqq k_{max} \leqq 2n-1 \tag{22}$$

[0105] Here, a residual section of $2^{2n}-(2\times 2_{max}^{k}-1)M$ values is $Z^c=Z^-\backslash_{i=1}^{M}P_i$, and is not selected for watermark embedding.

[0106] When an arbitrary code value x_1 belongs to the section P_i, a difference from the center value R_1 of the section is $d_i=x_i-R_1$, $x_i\in P_1$. Here, based on the range of $|d_i|$, the number k_1 of bits to be embedded in x_1 is determined as follows.

$$\sum_{l=0}^{k_i-1} 2^n < |d_i| \leqq \sum_{l=0}^{k_f} 2^n,\ k_i \geqq 1, \text{ if } x_i \neq R_1 \tag{23}$$

[0107] $k_i=0$, if $x_i=R_1$

[0108] Next, k_1 bits $\{w_I\}_{I=1}^{k_f}$ are embedded in x_1 as follows.

$$x_i' = R_i + 2_i^k d_i + \alpha(k_i) \text{ where } \alpha(k_i) = \text{sgn}(d_i)\sum_{l=1}^{k_f} 2^{t-1}w_1, \tag{24}$$

[0109] $x'_i \notin Z^t$ and $x'_{i-1}(n-1,n)\|x'_i(1,2) \notin Z^t$

[0110] The value $x_i=R_i$ which is the center value R_i of the section is the number of embedded bits $k_i=0$, and is excluded from bit embedding. Here, when a shifted code value x'_i is in the false start codon table Z^t or when the false start codon occurs between the x'_1 and the previous shifted code value x'_1, the number k_1 of embedded bits is reduced by one until reaching zero. This process is repeated. Thus, the false start codon is prevented in the same manner as a successive code value pair DE method. In this way, for all code values in the embedding target region, multiple bits are embedded depending on the number of embedded bits for each code value, and then the watermarked non-coding region Γ'(n) is obtained.

[0111] As additional information for watermark extraction and original sequence restoration, the number $K=\{k_i\}$ of embedded bits for each code value, a marker $T=\{\tau\}$ of a section shifted based on a section reference value and the LSB bit E of the 2-bit base binary number in the watermarked non-coding region Γ'(n) are required. Like the successive code value pair DE method, a bit string C of the additional information (K,T,E) is generated with lossless compression, and then the bit string is substituted by the LSB bit of the base binary number in Γ'(n). The DNA sequence $D'=D'^{nc}+D^c$, $D'^{nc}=\Gamma''(n)+\Gamma^c(n)$ containing the final additional information and the non-coding region Γ''(n) where the watermark is embedded is transmitted.

[0112] FIG. **7** shows code value shifting based on the difference |d| from the center value R_1 and a watermark bit when the maximum number of shifting bits on P_i is $k_{max}=3$. An arbitrary section P_i of a histogram domain is divided into a left subsection P_i^- and a right subsection P_i^+ based on the center value R_i. In the case of |d|=1, 3-bit (k=3) embedding is possible. In the case of $|d|\in\{2,3\}$, 2-bit (k=2) embedding is possible, and in the case of $|d|\in\{4,5,6,7\}$,1-bit (k=1) embedding is possible. In the case of |d|=0 and $x=R_i$, a bit is not embedded (k=0).

[0113] The code value x corresponding to the right subsection P_i^+ (d>0) of the section P_i is shifted by the watermark bit to the left subsection P_{i+1}^-(d≤0) of the right section P_{i+1}. In contrast, x corresponding to the left subsection P_i^-(d<0) of the section P_i is shifted by the watermark bit to the right subsection P_{i-1}^+(d>=) of the left section P_{i-1}. In other words, as shown in FIG. **8**A, the code value of the right subsection of the section P_i and the code value of the left subsection of the right adjacent P_{i+1} are shifted to each other. In contrast, the code value of the left subsection of the section P_i and the code value of the right subsection of the left adjacent P_{i-1} are shifted to each other.

[0114] Among the watermarked code values, the code value which is the center value $x'_i=R_i$ is generated in three cases. First, when the previous code value is the center value $x_i=R_i$ ($k_i=0$), it is excluded in shifting. Thus, the original code value $x_i=R_i$ is not shifted. Also, as shown in FIG. **8**A, the case is that values in the right subsection P_{i-1}^+ of the left section and in the left subsection P_{i+1}^- of the right section are shifted. The case where shifting is performed and the case where shifting is not performed can be distinguished by the number of embedded bits for each code value. Thus, for

extraction and restoration, the shifted previous section information T=$\{\tau\}$ is required as follows.

$$\tau = \begin{cases} 0, \text{if } x' = R_i \text{ and } x \in P_{i-1}^{+} \\ 1, \text{if } x' = R_i \text{ and } x \in P_{i+1}^{-} \end{cases} \quad (25)$$

[0115] As shown in FIG. **8**B, among M sections, code values from the right subsection P_1^+ of P_1 to the left subsection P_M^+ of P_M are shifted. Code values corresponding to the remaining boundary sections P_1^- and P_M^+ are assigned with the number of shifting bits k=0.

[0116] (2) Decoding and Restoration Processes

[0117] In decoding process of the present invention, from the non-coding region Γ''(n) of the DNA sequence D' previously transmitted, the additional information (K,T,E) of the compressed bit string is obtained, and then the watermarked non-coding region Γ'(n) by base binary number substitution of E is obtained. From the code sequence X' of Γ'(n) watermarking and original value restoration are performed by the number K of shifting bits for each code value and the marker of T=$\{\tau\}$ a shifted section.

[0118] When the code value x'_1 of the code sequence X^+ is given, the center value R of the original section of x'_1 is required to be obtained first. That is, when the shifted section P_1 of x'_1 is not the boundary section ($x'_i \in P_1$) and the number k_1 of shifting bits is $k_i>0$, the center value R for the previous section of x'_i is obtained as follows.

$$R = \begin{cases} R_{j-1}, \text{if } x'_i \in P_i^- \text{ or } (x'_i = R_j \text{ and } \tau_i = 0) \\ R_{j+1}, \text{if } x'_i \in P_i^+ \text{ or } (x'_i = R_i \text{ and } \tau_i = 1), \text{if } x'_i \in P_i \text{ and } k_i > 0 \end{cases} \quad (26)$$

[0119] Here, based on the shifted section P_i of x'_i, the center value R of the section before embedding is easily obtained. However, when x'_i is the center value R_i of the shifted region P_i ($x'_i=R_i$), $\hbar$ is obtained by the marker τ_i of the previous section. The watermark k_i bits $\{w_I\}_{I=1}^{k_i}$ on x'_1 and the original code value x_1 are obtained using the center value R of the previous section as follows.

$$w_I=((x'_i-R)>>(l-1))\%2 \text{ for } l=1, \ldots, k_i \quad (27)$$

$$x_i=R+((x'_i-R)>>k_i) \quad (28)$$

[0120] (3) Watermark Capacity and Additional Information

[0121] When the coding order n and the maximum number k_{max} of section shifting bits are given, the number of watermark bits embedded in the embedding region

$$\Gamma(n) = \{D_i\}_{i=1}^{|\Gamma(n)|}$$

is determined based on the number of bits defined by the difference range from the center value in the histogram domain section P_i and the frequency at which the code value belongs to each section.

[0122] The frequency with z value on the code value histogram is designated by p(z). Here, the number of shifting bits on an arbitrary section P_i is calculated by the sum of the number $C(P_i^-)$ of shifting bits in the left subsection P_i^- and the number $C(P_i^+)$ of shifting bits in the right subsection P_i^+.

$$C(P_j^+) = \sum_{i=0}^{k_{max}-1} \left(\sum_{t=0}^{2^i-1} p(R_j + 2^i + t)(k_{max} - i) \right), \text{ for } d > 0 \quad (29)$$

$$C(P_j^-) = \sum_{i=0}^{k_{max}-1} \left(\sum_{t=0}^{2^i-1} p(R_j - 2^i - t)(k_{max} - i) \right), \text{ for } d < 0 \quad (30)$$

[0123] The total number of watermark bits embedded in $\Gamma(n)=\{D_i\}_{i=1}^{|\Gamma'(n)|}$ is the sum of the number of shifting bits on the remaining sections, except for the boundary sections P_1^- and P_M^+ among total M sections, and the number of bits per base bpn $bpn_{HS}(n,k_{max})$ is defined as follows.

$$bpn_{HS}(n, k_{max}) = \frac{1}{\sum_{i=1}^{|\Gamma(n)|} N_i} \left(C(P_1^+) + \sum_{j=2}^{M-1} (C(P_j^+) + C(P_j^-)) + C(P_M^-) \right) \text{[bit/base]} \quad (31)$$

[0124] $|\Gamma(n)|$ is the number of embedding regions, N is the number of code values in the region D_i, and

$$\sum_{i=1}^{|\Gamma(n)|} N_1$$

is the total number of bases in the embedding target region.

[0125] The additional information $Extra_{HS}(n,k_{max})$ for watermark extraction and restoration is the number R of shifting bits for each code value, the marker T of the section shifted based on the section reference value, and the LSB bit E of the 2-bit base binary number of the watermarked non-coding region Γ'(n). When the maximum number of shifting bits in the histogram domain section is k_{max}, the number of embedded bits is expressed by $\lceil \log_2 k_{ma} \rceil$ bit. Thus, the number K of shifting bits for whole code values is expressed by total

$$\lceil \log_2 k_{max} \rceil \sum_{i=1}^{|\Gamma(n)|} N_1$$

bits. The marker T of the shifted section is binary information determining whether the code value $x'=R_i$ shifted based on the center value of the adjacent section is shifted from the left section or the right section, and is expressed by

$$T = \sum_{i=1}^{|\Gamma(n)|} N_i \times \sum_{i=1}^{M} p(x' = R_i)$$

bits. E is

$$\sum_{i=1}^{|\Gamma(n)|} |D_i|$$

bits that is the same as the number of bases of all regions in Γ'(n). Thus, additional information $Extra_{HS}(n,k_{max})$ is as follows.

$$\begin{aligned} Extra_{HS}(n, k_{max}) &= K + T + B \\ &= \lceil \log_2 k_{max} \rceil \sum_{i=1}^{|\Gamma(n)|} N_i + \sum_{i=1}^{|\Gamma(n)|} N_i \times \\ &\quad \sum_{i=1}^{M} p(x' = R_j) + \sum_{i=1}^{|\Gamma(n)|} |D_i| \\ &= \sum_{i=1}^{|\Gamma(n)|} \left(N_i \left(\lceil \log_2 k_{max} \rceil + \sum_{i=1}^{M} p(x' = R_i) \right) + |D_i| \right) [\text{bit}] \end{aligned} \tag{32}$$

[0126] When a compression rate is ρ, lossless compression is performed such that additional information $Extra_{HS}(n,k_{max})$

$$\rho \times Extra_{HS}(n, k_{max}) < \Phi \leqq \sum_{i=1}^{|\Gamma(n)|} |D_i|.$$

[0127] When the watermark bit is not embedded k=0, it corresponds to the boundary section of the histogram domain section, the residual section that do not belong to the section, and the code value that is the center value of the section. That is, k=0 probability P(k=0|x) is as follows.

$$P(k=0 \mid x) = \sum_{t=0}^{R_1-1} p(x=t) + \sum_{t=R_N+1}^{R_N+2^k \max -1} p(x=t) + \sum_{t=R_N+2^k \max}^{\square} p(x=t) + \sum_{j=1}^{M} p(x=R_j) \sum_{t=0}^{R_1-1} p(t)$$

is the probability of the code value in P_1^- section,

$$\sum_{t=R_N+1}^{R_N+2^{k max}-1} p(t)$$

is the probability of the code value in P_M^+ section, and

$$\sum_{t=R_N+2^k max}^{2^{2n}-1} p(t)$$

is the probability of the value in the residual section that do not belong to P. Last,

$$\sum_{i=1}^{M} p(R_j)$$

is the probability of the code values that are the center values of all sections.

$$P(k-1 \mid x), P(k=2 \mid x), \cdots P(k=k_{max} \mid x)$$

$$\sum_{i=0}^{k_{max}} P(k=i \mid x) = 1$$

[0128] 4.2 Circular Histogram Shifting (CHS)

[0129] Unlike the pixel value of the image, code values in the non-coding region have no condition for definition, and thus shifting between the maximum value and the minimum value is possible. In the circular histogram shifting method, histogram section shifting is changed to circular histogram shifting such that embedding is possible in the left subsection P_1^{-1} (d<0) of P_1 and in the right subsection P_M^+ (d>0) of P_M that are the boundary sections, thereby increasing watermark capacity in the non-circular histogram shifting method.

[0130] (1) Coding Process

[0131] In the rest sections except for the boundary sections and the residual section, the watermark is embedded in the same manner as embedding process of the non-circular histogram shifting method. In circular form of the histogram domain section, as shown in FIG. **9**, P_1^- and P_M^+ subsections, which are two boundary sections, are not shifted by the residual section. Thus, in the present invention, P_M^+ is shifted to the residual section such that two subsections of P_M are separated. That is, when the number of the code values in the residual section is $\delta=2^{2n}-(2\times 2_{max}{}^k-1)M$, P_M region is,

$$P_M = P_M^- + P_M^+ \tag{33}$$

[0132] where $P_M^- = \{z-2^{k_{max}}+1, \ldots, z-1, z\}$, $R_M^- = z$

[0133] $P_M^+ = \{z+\delta, z+\delta+1, \ldots, z+\delta+2^{k_{max}}-1(=2^{2n}-1)\}$, $R_M^{+=z+\delta}$,

[0134] divided into a subsection P_M^- smaller than $R_M^- = z$ and a subsection P_M^+ larger than $R_M^+ = z+\delta$. In P_M section, two center reference values are generated.

[0135] By the center value $\hbar$ of the section P_1 to which x_1 belongs on the arbitrary code value x_1

$$R = \begin{cases} R_j, \text{ if } x_i \in P_i \text{ for } j = 1, 2, \ldots, M-1 \\ R_M^-, \text{ if } x_i \in P_M^- \text{ for } j = M \\ R_M^+, \text{ if } x_i \in P_M^+ \text{ for } j = M \end{cases}, \tag{34}$$

[0136] k_1 bits $\{w_n\}_{n=1}^{k_f}$ are embedded as follows.

$$x'_i=(R+2_i^k d_i+\alpha(k_i))\%2^{2n} \quad (16)$$

[0137] where $d_i=x_i-R$ and

$$\alpha(k_i)=\mathrm{sgn}(d_i)\sum_{l=1}^{k_i}2^{l-1}w_l$$

[0138] Here, the number of shifting bits of the residual value $[R_M^-+1,R_M^+-1]$ between P_M^- and P_M^+ and the code values that are the center values of respective sections is zero.

[0139] Information T on the previous section for the value x'_1 shifted to the center value of the adjacent section is determined as follows.

$$\tau=\begin{cases}0, \text{ if } (x'=R_j \text{ and } x\in P_{j-1}) \text{ or } (x'=R_M^+ \text{ and } x\in P_1)\\ 1, \text{ if } (x'=R_i \text{ and } x\in P_{i+1}) \text{ or } (x'=R_1 \text{ and } x\in P_M^+)\end{cases} \quad (36)$$

[0140] In this way, watermarks are embedded into all code values in the code sequence X without occurrence of intra code and inter code false start codon, and the watermarked non-coding region Γ'(n) is obtained. The additional information required for watermark decoding and restoration of the original code value is the number K of shifting bits for each code value, the marker T of the shifted section, and the LSB bit E of a 2-bit base binary number, like the non-circular method. LSB substitution of the compressed additional information is applied in the same manner as the two methods, and the final watermarked DNA sequence D' by the substituted region Γ"(n) is transmitted.

[0141] (2) Decoding and Restoration Processes

[0142] Form the substituted region Γ"(n) of the transmitted DNA sequence, the watermarked region Γ'(n) is obtained by inverse substitution, and then from the code sequence X' in Γ'(n), the watermark is decoded by (K,T) and the original code sequence is restored.

[0143] When the code value x'_1 with $k_i>0$ is provided in the code sequence X', the center value R of the previous section of x'_1 is obtained depending on the boundary section and the non-boundary section as follows.

$$R=\begin{cases}R_{j-1}, \text{ if } x'_i\in P_j^- \text{ or } (x'_i=R_j \text{ and } \tau_i=0)\\ R_{j+1}, \text{ if } x_i\in P_i^+ \text{ or } x'_i=R_i \text{ and } \tau_i=1 \text{ for non-boundary region}\end{cases} \quad (37)$$

$$R=\begin{cases}R_M^+, \text{ if } 0\leqq x'_i<R_1 \text{ or } x'_i=R_1 \text{ and } \tau_i=0\\ R_1, \text{ if } R_M^+<x'_i\leqq 2^{2n-1} \text{ or } x'_i=R_M^+ \text{ and } b_i=1\\ \text{for boundary region}\end{cases} \quad (38)$$

[0144] k_1 bits $\{w_I\}_{I=1}^{k_f}$ and the original code value x_i are obtained by R as follows.

$$w_I=(((x'_i-R)\%2^{2n})>>(l-1))\%2 \text{ for } l=1,\ldots,k_i \quad (39)$$

$$x_i=R+((x'_i-R)\%2^{2n}>>k_i) \quad (40)$$

[0145] (3) Watermark Capacity and Additional Information

[0146] In the circular histogram shifting method, the watermark is embedded in all sections except for the residual section in the code value histogram domain range. Thus, when the coding order and the maximum number k_{max} of section shifting bits are given, the number of watermark bits in the embedding region Γ(n) is the sum of the number of shifting bits on the left subsection P_i^- (d<0) and the right subsection P_i^+ (d>0) of each section, and bpn $bpn_{CHS}(n, k_{max})$ thereof is as follows.

$$bpn_{CHS}(n,k_{max})=\frac{1}{\sum_{i=1}^{|\Gamma(n)|}N_i}\sum_{j=1}^{M}(C(P_j^+)+C(P_j^-))\ [\text{bit}] \quad (41)$$

[0147] The additional information $Extra_{HS}(n,k_{max})$ for watermark extraction and restoration is the same as information in the non-circular histogram shifting method, $Extra_{HS}(n,k_{max})=Extra_{CHS}(n,k_{max})$. Like the above-described methods, lossless compression is performed such that the additional information $Extra_{CHS}(n,k_{max})$ is

$$\rho\times Extra_{CHS}(n,k_{max})<\Phi\leqq\sum_{i=1}^{|\Gamma(n)|}|D_i|.$$

The circular histogram shifting method has the same additional information but higher watermark capacity, compared to the non-circular histogram shifting method.

[0148] The previous region information of the code value shifted to the center value and information on the number of embedded bits of the code value that belong to all regions except for the residual value region are follows.

$$N_E^{CHS}=N\times\left[p(x'\epsilon R)+\left(1-\sum_{t=R_N+1}^{R_N-1}p(t)\right)\times\lceil\log_2 k_{\max}\rceil\right][\text{bit}] \quad (42)$$

[0149] Here,

$$\sum_{t=R_1+1}^{R_N-1}p(t)$$

is probability of belonging to the residual value, and $\hbar$ is reference value $R=\{R_1, R_2, \ldots, R_{M-1}, R_{M1}, R_{M2}\}$ of the region. Thus, the bpn of additional data is $bpn_E^{CHS}=N_E^{CH}/N_D$ [bit/base]. Capacity efficiency O^{CHS} that is a ratio of additional data to the embedded data is $C^{CHS}=N_W^{CHS}/N_E^{CHS}=bpn_W^{CHS}/bpn_E^{CHS}$.

[0150] Although a preferred embodiment of the present invention has been described for illustrative purposes, those skilled in the art will appreciate that various modifications, additions and substitutions are possible, without departing from the scope and spirit of the invention as disclosed in the accompanying claims.

What is claimed is:

1. A reversible DNA information hiding method based on prediction-error expansion and histogram shifting, the method comprising:

coding, at a first step, a four-letter base sequence of a non-coding region DNA to an n order code value;

embedding, at a second step, multiple bits for each code value by a least square (LS) prediction error;

embedding, at a third step, an n order watermark bit by non-circular histogram and circular histogram multi-level shifting;

verifying, at a fourth step, occurrence of a start code of a watermarked intra code value and a watermarked inter code value.

2. The method of claim **1**, wherein at the first step,

b is a four-letter base b={'A', 'T', 'C', 'G'}, b is a base value of the b, x is a base block consisting of n bases, x is a code value for the base block x, and n is a coding order,

coding to a 2n-bit code value x in units of the base block x consisting of the n bases is performed as follows

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

where $x=(b_1, b_2, \ldots, b_n)$, $x \in [0, 2^{2n}-1]$ and

The bases of the base block are restored from the code value x as follows

$f^{-1}(x)=x$ where $b_k=(x>>2(n-k))\%4$ for $k=1, \ldots, n$.

3. The method of claim **1**, wherein at the fourth step, preventing of a false start codon in the watermarked intra code value comprises:

generating a code value table containing the false start codon in advance; and

embedding a watermarked code value not to contained in the code value table.

4. The method of claim **1**, wherein at the fourth step, preventing of a false start codon in the watermarked intra code value comprises:

when a previous watermarked code value x'_{1-1} is given, a number of embedded bits for a current processed code value x'_1 is controlled such that the current processed code value x'_1 does not satisfy

$$x'_{1-1}(n-1,n)\|x'_1(1,2) \in Z^c$$

if $(x'_{1-1}\%2^4)=f('AT')=1$ and $(x'_1>>2(n-1))\%2^2=f('G')=3$

if $(x'_{1-1}\%2^2)=f('A')=0$ and $(x'_1>>2(n-2))\%2^4=f('YG')=7$.

5. The method of claim **1**, wherein at the second step, the code value is predicted through local prediction for each embedding region.

* * * * *