An improved Huffman coding method for archiving text, images, and music characters in DNA

Menachem Ailenberg and Ori D. Rotstein
Departments of Surgery, University of Toronto, and St. Michael’s Hospital, Li Ka Shing Knowledge Institute, Keenan Research Centre, Toronto, Ontario, Canada

BioTechniques 47:747-754 (September 2009) doi 10.2144/000113218
Keywords: DNA archiving; information storage in DNA

An improved Huffman coding method for information storage in DNA is described. The method entails the utilization of modified unambiguous base assignment that enables efficient coding of characters. A plasmid-based library with efficient and reliable information retrieval and assembly with uniquely designed primers is described. We illustrate our approach by synthesis of DNA that encodes text, images and music, which could easily be retrieved by DNA sequencing using the specific primers. The method is simple and lends itself to automated information retrieval.

Introduction
The increasing use of digital technology presents a challenge for existing storage capabilities. The need for a reliable and long-term solution for information storage is further heightened by the prediction that the current magnetic and optical storage will become unrecoverable within a century or less (1). DNA is a compact, long-term, and proven medium for information storage. Indeed, over the last few decades, a good case has been made for crucial information storage in DNA (2). Desirable properties of DNA include its capacity for long-term information storage and recovery, which are mostly independent of technological changes, the ability to conceal data in a miniaturized fashion and its ability to be transferred, when required, via self propagation (1–6). Various approaches for information coding in DNA have been reported, including the Huffman code, the comma code, and the alternating code (4), a straight coding based on 3 bases per letter (1,2,6), or sequential conversion of text to keyboard scan codes followed by conversion to hexadecimal code and then conversion to binary code with a designed nucleotide encryption key (5). Each approach offers advantages and inherent difficulties, and differs in the degree of economical use of nucleotides. We sought to develop an alternate approach for information archiving in DNA. We used the principles of the Huffman code (4,7) to define DNA codons for the entire keyboard, for unambiguous information coding. The approach described in this manuscript is based on the construction of plasmid library for information archiving with specially designed primers embedded in the message segment with an exon/intron structure for rapid, reliable, and efficient information retrieval.

Materials and methods
The DNA coding was based on modification of the Huffman code (2,4,7,8). We also adopted the nomenclature suggested by Cox (2) for definition of the DNA segment representing a single character as ‘codon’. DNA (844 bp; Figure 1A) was synthesized and inserted as a Stel/KpnI fragment in pBluescript-based plasmid (Mr. Gene GmbH, Regensburg, Germany). Sequence confirmation of supplied plasmid was provided by the manufacturer using plasmid universal primer. For information retrieval, plasmid (300 ng/7 μL) was mixed with sequencing primer (5 pmol/0.7μL; Sigma, Oakville, Ontario) (Figure 1B) and subjected to sequencing (service was performed by The Centre for Applied Genomics, The Hospital for Sick Children, Toronto, ON, Canada). The chromatogram was created using the FinchTV 1.4 application (Geospiza Inc., Seattle, WA, USA). Sequences of designed and sequenced DNA were aligned using blast2seq (NCBI, Bethesda, MD, USA). PCR amplification was performed in iQ5 cycler (Bio-Rad Laboratories, Mississauga, ON, Canada). A reaction mixture contained 2 units Taq polymerase with 1x reaction buffer (New England BioLabs, Pickering, ON, Canada), 0.2 mM each dNTP (Fermentas, Burlington, ON, Canada), 0.3 mM each primer, 200 ng plasmid DNA, and UltraPure distilled water (Invitrogen, Burlington, ON, Canada) to a volume of 20 μL. PCR conditions were 94°C for 3 min; 94°C for 30 s, 55°C for 30 s, and 72°C for 60 s for 30 cycles; then 72°C for 7 min for final extension; and held at 4°C. Ten microliters of PCR reaction was mixed with 2 μL 6× loading buffer (Fermentas). DNA fragment size was determined by loading in parallel 5 μL 100-bp DNA ladder (Fermentas) and resolved on 1% agarose gel (BioShop, Burlington, ON, Canada). Gel was visualized with UV transillumination, and image was captured with Biospectrum AC Imaging System (UVI, Upland, CA, USA).

Results and discussion
Rationale for the improved Huffman Method
One of the prerequisites for a good DNA coding method is the economical use of nucleotides per character. In the improved Huffman coding described in this paper, the bases-to-character ratio was ~3.5. This number is more economical compared with previously described methods that enable entire keyboard coding (e.g., the comma or alternating codes (6 bases/character) (4) or sequential encryption (~5.3 bases/character) (5)). It should be noted that other coding methods yielded lower base-to-character ratios, but these approaches were limited to a low number of characters, usually sufficient only for text encoding of the English alphabet (1,2,6). Information storage of DNA in living organisms has a disadvantage of losing the information as a result of breakage by mutation, deletion, and insertion of DNA (9). Yachie et al. (5) described an alignment-based approach for prolonged and reliable information storage in DNA in living organisms. This approach provides efficient recovery of stored data even for damaged DNA (9). While the size of foreign DNA inserted into a living organism may be limited (2), recent studies have
A

**Figure 1. DNA sequence of insert and sequencing primers.** (A) Sequence of 844-bp DNA that stores information for the text, music, and image of the nursery rhyme Mary Had a Little Lamb described in the text (Figures 2–4 and 6; Supplementary Figures 1 and 2). The text’s sequence is shown in orange type, music in black type, and image in red type. Page break is shown in bold underlined style. Characters like punctuation, spaces, and “shift” keys are shown in lower case. Note that the sequence of the music is interrupted with sequencing primer 2 (intron). According to the assembly rules of the information sequence (exon), the sequence of the primer is recognized as an intron, excised, and the sequence of the flanking exons are spliced and read as continuous information. The 844-bp DNA segment was synthesized and inserted into a pBlueScript-based plasmid as a SacI/KpnI fragment. (B) Sequencing sense and anti-sense primers used for information retrieval. Blue bold type depicts the 5’ end of sense primers and corresponding 3’ sequence of anti-sense primers. Green bold type depicts the 3’ end of sense primers and corresponding 5’ sequence of anti-sense primers. Lowercase black type depicts seven bases of random sequence that flank the plasmid and primer number, designed to provide uniqueness to each primer and reduce mis-priming in the sequencing reaction. Pink bold type depicts the plasmid number (in this particular example, the number 1 in all cases). Red uppercase type depicts primer number (in this particular case, 1 and 2 for sense primers, and 0 and 2 for anti-sense primers). For this particular study, we assumed a 1-digest plasmid number. When more than 10 plasmids are present, a space should be inserted to distinguish between the plasmid and primer number. (C) DNA codons assigned to plasmid and primer number codes.

**General information about the library** is initially retrieved by DNA sequencing from the index plasmid using plasmid-specific universal primers and then from the plasmid library by uniquely designed primers (Figure 1B). Bancroft et al. (1) described a technique for information storage in DNA that utilized two classes of DNA: information-containing DNA and polyprimer key (PPK) DNA. According to this method, the PPK contained the entire sequences of the primers. In contrast, our index plasmid contains only the information for the structure of the information library, and the sequencing primers are embedded in the information library (for further description of information retrieval, see Supplementary Materials). The Huffman code for DNA encryption suggested by Smith et al. enables encoding for only 26 characters (4). We sought to improve this approach in order to enable coding of the entire standard computer keyboard. Since the retrieving primers in our method contain GC bases in the 5’ and 3’ ends, and the codons suggested by Smith et al. are GC-rich, we first replaced Cs with As and Gs with Ts, and moved, when possible, the GC-containing codons down the frequency table (see below).

**Rules for coding of text, music, and images with the improved Huffman coding**

We defined rules for text, music, and image coding. For text coding (coding preceded with “tx” GTG TTCT TACCA), we created three columns headed by low-base number DNA codons (G,T,T,T), and placed the remaining codons in increasing base number under each header codons (Figure 2; Supplementary Materials). For musical notes coding (coding preceded with “mu” TAAC TACT TACCA), we utilized the one-column—modified Huffman coding (Figure 3; Supplementary Materials). For image coding (coding preceded with “im” GCG TAAC TACCA), we utilized the one-column—modified Huffman coding (Figure 4; Supplementary Materials).

According to our design, the sequence encoding the text, music, and image is part of the information sequence. Sequencing primers are embedded in the information sequence in 500-bp intervals (Figure 5A). This structure is not unlike genomic exon/intron structure of genes. In our case, the exons are usually 500 bp, and the introns (sequencing primers) are 20–30 bp in size.

This unique repetitive structure allows for easy pattern identification, even for those who are not familiar with the rules of our specific coding. This pattern also allows for an algorithm to decipher the reading frame of the message.

**The index plasmid.** The index plasmid is designed to contain an insert with general identification information (Figure 5B). The index plasmid is different from the library information plasmids. Therefore, for storage purposes, the index plasmid can be mixed with the library plasmids, and the index information can be retrieved with generic plasmid-specific primers. For more information on the index plasmid, see the Supplementary Materials. In addition, the index plasmid contains sequences of clusters of 14 (7 + 7) bases representing the random sequences of the first primer in each plasmid. This is essential for the initial sequencing of the plasmid and retrieving the remaining sequencing primers, if they are not available. Thus, in our particular example, the index plasmid also contains at its 5’ end the sequence CGGTGTAC-GAC ACT. For the purpose of simplicity, we describe only the theoretical aspects of...
Illustration of the improved Huffman coding

To illustrate our method, a 844-bp DNA fragment (Figure 1A) was synthesized. This DNA fragment contained encoded information for text, musical notes, and image for the nursery rhyme “Mary Had a Little Lamb,” by Sarah Josepha Hale (10), and was constructed according to the principles outlined in Figures 1–5. The DNA sequence for the text and musical notes of the rhyme are shown in Figures 2 and 3, respectively. We also show here a simple image of a lamb (Figure 6). The image is drawn in a field of 10 × 10 units. The head, body and ear of the lamb are defined by ellipses, the eye by a circle, the legs by lines, and the tail by a rectangle. It should be noted that the concise definition of the geometrical shapes in DNA codes described here allows for economical coding of the lamb image by only 238 bp of DNA. For retrieving information from the 844-bp DNA fragment in our particular example, sense primers 1 and 2 were flanking a 500-bp segment, and anti-sense primers 2 and 0 were flanking a 344-bp segment (Figure 1). To illustrate the utility of the specificity of the primer design, a PCR reaction employing the 3 primer sets was performed. As demonstrated in Figure 7, the 3 amplification products corresponding to the expected amplicon sizes were noted.

Unique sequencing primers for information retrieval

The sequencing primers (Figures 1 and 5) were flanked by 5′-GCC and 3′-GCC (sense) or 5′-GCG and 3′-GCG (anti-sense) for easy identification and also for creating the index plasmid without actual synthesis of the insert.

ILUSTRATION OF THE IMPROVED HUFFMAN CODING

To illustrate our method, a 844-bp DNA fragment (Figure 1A) was synthesized. This DNA fragment contained encoded information for text, musical notes, and image for the nursery rhyme “Mary Had a Little Lamb,” by Sarah Josepha Hale (10), and was constructed according to the principles outlined in Figures 1–5. The DNA sequence for the text and musical notes of the rhyme are shown in Figures 2 and 3, respectively. We also show here a simple image of a lamb (Figure 6). The image is drawn in a field of 10 × 10 units. The head, body and ear of the lamb are defined by ellipses, the eye by a circle, the legs by lines, and the tail by a rectangle. It should be noted that the concise definition of the geometrical shapes in DNA codes described here allows for economical coding of the lamb image by only 238 bp of DNA. For retrieving information from the 844-bp DNA fragment in our particular example, sense primers 1 and 2 were flanking a 500-bp segment, and anti-sense primers 2 and 0 were flanking a 344-bp segment (Figure 1). To illustrate the utility of the specificity of the primer design, a PCR reaction employing the 3 primer sets was performed. As demonstrated in Figure 7, the 3 amplification products corresponding to the expected amplicon sizes were noted.

Unique sequencing primers for information retrieval

The sequencing primers (Figures 1 and 5) were flanked by 5′-GCC and 3′-GCC (sense) or 5′-GCG and 3′-GCG (anti-sense) for easy identification and also for creating the index plasmid without actual synthesis of the insert.

ILLUSTRATION OF THE IMPROVED HUFFMAN CODING

To illustrate our method, a 844-bp DNA fragment (Figure 1A) was synthesized. This DNA fragment contained encoded information for text, musical notes, and image for the nursery rhyme “Mary Had a Little Lamb,” by Sarah Josepha Hale (10), and was constructed according to the principles outlined in Figures 1–5. The DNA sequence for the text and musical notes of the rhyme are shown in Figures 2 and 3, respectively. We also show here a simple image of a lamb (Figure 6). The image is drawn in a field of 10 × 10 units. The head, body and ear of the lamb are defined by ellipses, the eye by a circle, the legs by lines, and the tail by a rectangle. It should be noted that the concise definition of the geometrical shapes in DNA codes described here allows for economical coding of the lamb image by only 238 bp of DNA. For retrieving information from the 844-bp DNA fragment in our particular example, sense primers 1 and 2 were flanking a 500-bp segment, and anti-sense primers 2 and 0 were flanking a 344-bp segment (Figure 1). To illustrate the utility of the specificity of the primer design, a PCR reaction employing the 3 primer sets was performed. As demonstrated in Figure 7, the 3 amplification products corresponding to the expected amplicon sizes were noted.

Unique sequencing primers for information retrieval

The sequencing primers (Figures 1 and 5) were flanked by 5′-GCC and 3′-GCC (sense) or 5′-GCG and 3′-GCG (anti-sense) for easy identification and also for creating
a triple-GC clamp at the 3′ end for tight hybridization. In the middle of the primer, we reserved space for coding the plasmid number and the primer number. Primer number 1 indicates the sequence of the 5′ segment of the information insert, and primer number 0 indicates the sequence of the 3′ end of the information insert. Other primers are identified in ascending order, to a maximum of 20 in a 10,000-bp information insert. We specified here the plasmid and primer number in single digits. However, when more than 10 plasmids are included in the library, a space character (GAT) should be inserted to distinguish between the plasmid and primer numbers. Importantly, the plasmid and primer number encoded in the sequencing primer were designed to be flanked by a random seven-base sequence (a total of 14 bases per primer) to provide primer specificity when used for sequencing both in the sense and anti-sense orientation, or for PCR amplification.

Information retrieval by sequencing

In general, sequencing reactions do not provide sequencing information immediately.
Figure 6. Image of lamb encrypted in DNA. (A) DNA codons were defined as described in Figure 4. Upper row identifies the organ of the lamb, middle row defines the shapes and their parameters as defined in Figure 4, and lower row shows image encrypted in DNA codons. Also shown at the top of this panel is the DNA code that precedes the image (“im*”). (B) Image was created by using shapes (ellipse, circle, line, rectangular) as described in the “Results and discussion” section and Figure 4. A field of 10 x 10 units was created, and shapes were inserted. An example of the tail sequencing is shown in Supplementary Figure 2.

Figure 7. PCR amplification of DNA encrypting text, image, and music, using specific primers. PCR was performed using as template a plasmid containing the 844-bp synthesized DNA that encrypts text, music, and image. Numbers on the left indicate DNA fragment size, in bp. Lane 1, 100-bp DNA ladder; Lane 2, amplification using sense primer 1 and anti-sense primer 2; Lane 3, amplification using sense primer 2 and anti-sense primer 0; Lane 4, amplification using sense primer 1 and anti-sense primer 0. Note that the amplified bands in lanes 2, 3, and 4 correspond to the expected sizes of 553 bp, 314 bp, and 844 bp, respectively. For conditions of PCR amplification, consult the “Materials and Methods” section. For primer assignment, see Figure 1.

adjacent to the sequencing primer. Moreover, the sequence could occasionally be <500 bp, or a base call can be inconclusive. Therefore, sequencing in both orientations may be required. With the unique design of our primers, sequencing could be achieved with a high degree of accuracy. A good sequencing practice for retrieving information from an information library might be to initially use the sense primers only, and anti-sense primer 2 to retrieve the information of the 5’ end. Thereafter, the other anti-sense primers can be utilized as required. For sequencing in our particular example, we used sense primers 1 and 2, and anti-sense primers 2 and 0. Sequencing with primer 1 yielded a 1276-bp product. Since our insert was 844 bp, it also yielded a plasmid sequence downstream from the insert. However, the sequence was missing 41 bp just downstream of primer 1. This sequence was retrieved with anti-sense primer 2. Together, these two sequencing reactions retrieved the original information with 100% accuracy. We recognize that a sequencing reaction cannot always retrieve 1276 bp as was the case with primer 1, and therefore acknowledge that additional sequencing may be required. As mentioned above, we further retrieved the insert information with additional sequencing with sense primer 2 and anti-sense primer 0, again with 100% accuracy, compared with the initial designed insert. We then decoded the information obtained by DNA sequencing with the guidelines provided in Figures 2-4, and were able to reconstruct the text, music, and image encrypted in the DNA. An example of part of a sequencing chromatogram achieved with sense primer 1 is shown in Supplementary Figure 2. This particular sequenced information is the DNA sequence coding the rectangle that constructs the image of the lamb’s tail (bases 647-674 on the chromatogram) with 100% accuracy, compared with the original sequence.

In addition to the inherent advantageous attributes of information storage in DNA, our improved Huffman coding method for use of unambiguous DNA coding for archiving offers economical, easy pattern recognition and message retrieval through specially designed primers. As DNA synthesis and sequencing become faster and cheaper (Genome Analyzer Sequencing System, Illumina, San Diego, CA, USA; 454 Sequencing, Roche, Branford, CT, USA) information storage in DNA becomes even more attractive.

Acknowledgments

This study was supported by the Canadian Institutes of Health Research (CIHR; grant no. 37779). The authors declare no competing interests.

References


Received 28 March 2009; accepted 9 July 2009.

Address correspondence to Menachem Ailenberg, St. Michael’s Hospital, 30 Bond Street, 16CC-044, Toronto, Ontario, Canada M5B 1W8. email: m.ailenberg@uttoronto.ca