

Supplementary Material For:

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

Supplementary Methods

Text, music and image coding

Text coding. Frequencies of characters were determined by scanning scientific papers for frequency of the characters of a standard computer keyboard. The frequency of each character was, in principle, the same as that described for text frequencies (4, 11) except for the letters x, and z, which—although exhibiting low frequencies—were somewhat more frequent in scientific text than in literature text. The most frequent characters (e.g., “Space” key, “Shift” key, the letter e) were placed in the first column under base “G” heading. Thus, the DNA code for the letter “e” is GCT (G is the header of group 1 code and CT is the letter “e” code), and the coding for the number “3” is TTAAT (TT is the header of group 2 code and AAT the number “3” code). Since our method allows for many characters, a specific code was assigned to characters that are otherwise determined using the “Shift” key. For example, the “&” character is coded in our method with the six bases TTAAAA. Using the keyboard configuration, this character would be coded as “Shift” and “7” that would be otherwise coded with the 8 bases GTCTTCAA (GTC, “Shift,” TTCAA, “7”). A unique code for “Page break” was defined as “Shift” and “Tab” (GTCTACCC).

Music coding. We defined the seven musical notes A–G (Figure 3). For simplicity, we used only one octave, although varied octaves can be easily defined by assigning different numbers to different octaves. We also defined note values [whole note to sixteenth note, (see Figure S1), meters (2/4, 3/4, 4/4), and repeat (when a sequence is repeated more than one time)]. When defining musical notes that require more characters, the three-column Huffman coding can be used.

Image coding. We assigned DNA codons to various geometrical shapes (e.g., circle “C” is defined by “CAC”). The shape is followed by size and location values. For example, C r;x1;y1 represents a circle with radius of r units and its center coordinates are x1 and y1. We have also used this approach to draw other shapes utilizing polynomial trendline (data not shown). The coding principles of polynomial function are shown in Table 3. Again, for more complex images, the 3-column coding, similar to text coding, can be utilized. We have also designed principles for image animation, denoted by “an*” (GAC TTAT TACCA). This entails translocation of geometrical shapes from one set of coordinates (as defined in Figure 4) to another, with the inclusion of a time “t” parameter for specifying the rate of relocation. For simplicity, we did not include any further animation data in this communication. It should be noted that when more characters are required, more columns (e.g., a fourth column with AT header) could be added. In that case, the other codons should be reassigned. For example, the “Space” character defined in the three-header Figure 2 as “GAT” would be assigned as “GCT” and the other characters would be reassigned accordingly.

The index plasmid

In our specific example, the index plasmid contains information of the title (“DNA archive”), date (“1.1.2008”), authors (“Ailenberg, Rotstein, et al.”), affiliation (“SMH”), number of total pages (“p3”) and number of plasmids in the library (“p11”). The DNA sequence for this information is:

TTAC TTAT GAC GAC TTTG
TTCG TTTC GCG GCAT GCT tate
GCCG TAAT GCCG TAAT GCCC

GCCT GCCT TTCAC tate gtcGAC
GCG TTAG GCT TTAT GAAG GCT
TTTG GAAT TATC gtcTTTG GAG
GTG TTCT GTG GCT GCG TTAT gat
GCT GTG gat GAC TTAG tate GAAC
TTAAT tate GAAC TTAG GCCG.

Information retrieval

Direct sequencing. The plasmid library is prepared in aliquots of a lyophilized mixture with each plasmid in sufficient concentrations for direct sequencing after reconstitution (30–70 ng/μL each, according to the number of plasmids in the library). The first primer from each plasmid is constructed based on information retrieved from the index plasmid (alternatively, the first primer from each plasmid can be synthesized and stored separately). For information retrieval, sequencing reactions are set each with the different primer representing a different plasmid. The information of the remaining primers is collected by primer walking through the inserts, and used for subsequent sequencing.

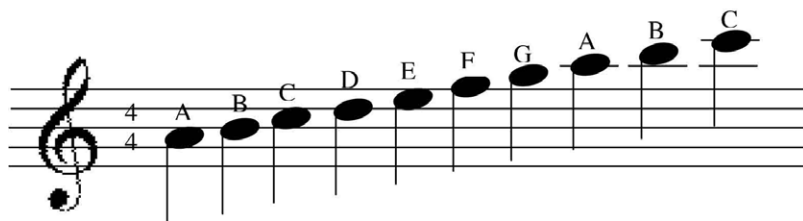
Sequencing of PCR amplicons. The plasmid library is prepared as described (in the “Direct sequencing” section), but in lower concentration. In this case, the first primer 1 and last primer 0 of each plasmid are also archived. A PCR reaction is run using sense and anti-sense primers for each plasmid, flanking the entire insert. The amplified product (10,000 bp) is purified and subjected to sequencing, using the specific primers from within the sequence. It should be noted that PCR amplification of relatively large pieces of DNA (up to 50 kb) is employed routinely and is commercially available.

Sequencing of DNA archived information with known primer sequence. In this case, the primer sequences are stored as a hard copy on conventional storage such as computer or paper. This approach is useful when a large body of information needs to be compressed; the primers for direct sequencing or PCR (described above) are selected according to the segment of the information to be retrieved. The location of the information in the plasmid library is determined by sequencing the index plasmid.

A

Name	Note
Whole Note	♩
Half Note	♪
Quarter Note	♩
Eighth Note	♪
Sixteenth Note	♪

B



C

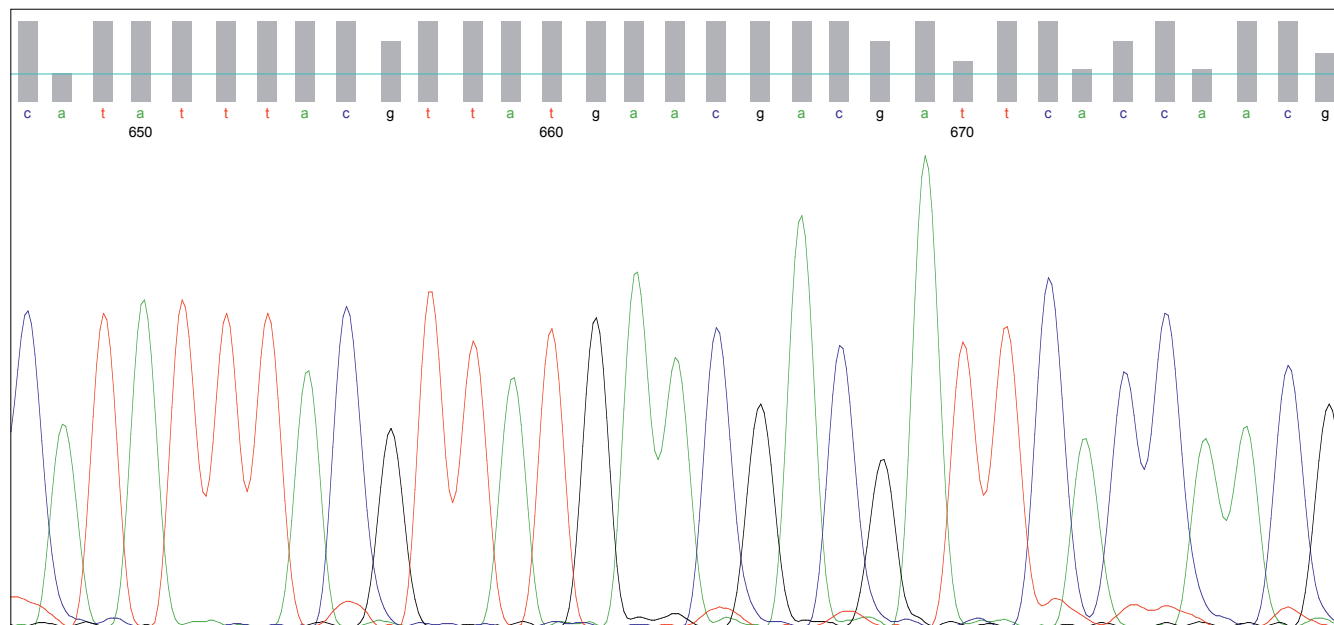
Ma-ry had a lit-tle lamb, lit-tle lamb, lit-tle lamb
 Ma-ry had a lit-tle lamb. Its fleece was white as snow.
 Eve-ry where that Ma-ry went, Ma-ry went, Ma-ry went,
 Eve-ry where that Ma-ry went, the lamb was sure to go. It
 fol-lowed her to school one day, school one day, school one day. It
 fol-lowed her to school one day, Which was a-gainst the rules. It
 made the chil-dren laugh and play, laugh and play, laugh and play. It
 made the chil-dren laugh and play, To see a lamb in school.

©EnchantedLearning.com

Figure S1. Musical notes for the nursery rhyme *Mary Had a Little Lamb*. (A) Note values. (B) Musical notes. Also shown on the left 4/4 m. (C) Musical notes and lyrics of the rhyme are from Enchanted Learning (2003, www.EnchantedLearning.com).

Finch Server File: MA-Mary1-17a

 Sample Name: 63690
 Mobility: KB_3730_POP7_BDTv3.mob
 Spacing: 15.7948
 Comment: <ID:63690>

 Signal Strengths: A = 1170, C = 971, G = 978, T = 1447
 Lane/Cap#: 64
 Matrix: n/a
 Direction: Native


Printed: Feb 3, 2008 11:37AM

 FinchTV v.1.4.0

Page 21 of 42

Figure S2. Part of chromatogram showing DNA sequencing of the rectangle shape of the tail of the lamb depicted in Figure 6B. The rectangle is encoded by bases 647–674 on the chromatogram. The information plasmid was sequenced with sense primer 1. Note the 100% accuracy compared with the designed sequence depicted in bold type in Figure 6A. Chromatogram was created using FinchTV 1.4 (Geospiza Inc., Seattle, WA, USA).