Haar Wavelet Transform of The Signal Representation of DNA Sequences

MOHAMED EL-ZANATY [1], MAGDY SAEB [1], A. BAITH MOHAMED[1],
SHAWKAT K. GUIRGUIS[2]

1. School of Engineering, Computer Department,
Arab Academy for Science, Technology and Maritime Transport
Alexandria, EGYPT
2. Information Technology Department,
Institute of Graduate Studies & Research,
Alexandria University, EGYPT

***Abstract*:** Complex sequences of DNA nucleotides and their associated search techniques can be relatively simplified when presented as a digital signal. This approach applies known signal processing techniques for the analysis of genomic information. We present a set of tools for the signal representation and analysis of genomic information. In this work, we provide a matrix and a sparse polynomial representation of the DNA. We show that sparse polynomial representation of the DNA sequences improves the search performance and reduces the storage requirements. The DNA nucleotides are presented using the compact form similar to QR (Quick Response) representation that offers a broad scope of practical usage. In addition, we speed up the search process by applying the Haar Wavelet technique on the resulting DNA signals. Based on multi-level Haar transform, the search starts at the n-th trend and follow the various levels upward until a match is found. Some important features of nucleotide sequences are revealed using these visual signal representations by comparing members of the same evolutionary family.

***Key words*:** DNA, Signal Analysis, Haar Wavelet, Multi-resolution Analysis, Sparse Polynomial, Quick Response.

1. Introduction

DNA, over millions of years, has demonstrated its effectiveness as a coding medium for the instruction set that governs and propagates living things. DNA is an appealing media for data storage due to the very large amounts of data that can be stored in a compact volume. DNA storage capacity vastly exceeds the storage capacities of conventional electronic, magnetic and optical media. A gram of DNA contains about 10^{21} DNA bases, or about 10^{8} tera-bytes. Hence, a few grams of DNA may have the potential of storing all the data stored in the world [1]. Different methods of DNA database search were developed, but most of these methods were built on string matching and bases alignment, therefore they needs much processing time and results have no established standards [1,2,3]. For example in order to find this sequence in the database shown in Table 1, we need to move through 300021 rows and make string matching (base alignment) for each one.

"ATTCTTCG.....TAGTCGT"

This is very slow and consumes relatively large amounts of power. The complexity of this process depends on the table size n which means O (n). Parallel processing may speed up this process, however with added hardware and software.

BLAST (Basic Local Alignment Search Tool) is one of the most widely used bioinformatics programs. It addresses a fundamental problem and the algorithm emphasizes speed over sensitivity [7]. Before fast algorithms such as BLAST and FASTA (Fast All) were developed, performing database searches for the protein or nucleic sequences was very time consuming.

 Manuscript received June 6, revised June 12

ID	Sequence
1	GTAGTTAA……AGTGTGC
2	ATTCTTCG……TAGTCGT
3	TTCGTCGG……TGTCTGT
:	…….
:	…….
300021	ATTCTTCG……TAGTCGT
:	…….
:	…….
	ATGCTGCG……ATTTGGGG

Table 1
A representative DNA database

This is achieved by using a full alignment procedure like Smith-Waterman. Indeed, BLAST is faster than Smith-Waterman but it cannot "*guarantee the optimal alignments of the query and database sequences*", as Smith-Waterman does. Smith-Waterman "*ensured the best performance on accuracy and the most precise results*" at the expense of time and computer power consumption. BLAST is more time efficient than FASTA by searching only for the more significant patterns in the sequences, but with comparative sensitivity. We observe BLAST and FASTA built on searching only for the more significant patterns in the sequences not for exact matching of all the sequence. Therefore, they cannot be used in many practical problems which need exact matching like crime detection. Even the known exact matching techniques, such as Smith-Waterman, are very time and power consuming [7]. On the other hand, DNA signal representation will transfer the data stored in DNA from its biological space to the signal space. Hence, this proposed approach will take advantage of the techniques of signal processing. The signal representation of DNA sequences will enable us to apply wavelet transforms to the developed signal. Hence, we can extract some important information. The usual *random walk* is a stochastic process represented by a sequence of partial sums of random variables. A one-dimensional random walk is constituted of a sequence of independent displacements, either left or right at each time step. The *DNA walk* denotes a special case, where the partial sums are obtained by aggregating the numerical values associated with the components of a DNA sequence. The walk is created by defining an incremental variable that associates to the time step *k* the value x (k) = ±1, depending on whether the base at position *k* along the sequence is respectively, a pyrimidine (C, T) or a purine (A, G). The DNA walk is defined as:

$$F(k) = \sum_{n=1}^{k} x(n)$$

This equation represents the DNA sequence as a series of integers for example the DNA sequence

ATTCCGGAGCTAG

After applying the DNA walk, the string will take the form:

-1,0,1,2,3,2,1,0,-1,0,1,0,-1

A continuous one-dimensional Daubechies wavelet was applied to analyze the DNA walk to detect regions of irregular patterns. Based on the above literature review, one concludes that DNA signal representations are very accommodating in several applications [8, 9, 10].

In the next four sections we discuss different representations and the formulation of the problem. In section 2, we present the binary and quaternary representation of the DNA sequence. Section 3 discusses the matrix and sparse polynomial compact representation of the DNA sequence. In Section 4 the Haar wavelet is applied on the resulting signal and provides the foundation for speeding up the search process in the database. Section 4 deals with the matrix representation of the DNA sequence and the sparse polynomial representation. In this section we present the QR (Quick Response) representation of the DNA sequence as graphical representation. A comparison between human and chimp genomes based on multi-resolution Haar transform is also discussed. Finally, we provide a summary and our conclusions. The pseudo-code of the program utilized in this work is shown in the Appendix.

2. DNA Binary and Quaternary Representations

Mapping the DNA sequences to binary representation is a simple straight forward procedure. For most tasks, a flat encoding of 2 bits/nucleotide, assigned in alphabetical order would be a sufficient starting point.

$A = (00)_2$	or	$A = 0_Q$
$C = (01)_2$	or	$C = 1_Q$
$G = (10)_2$	or	$G = 2_Q$
$T = (11)_2$	or	$T = 3_Q$

For example the DNA sequence

ATTCCGGAGCTAG

It will be represented in binary format as:

$$(0011110101101000100111001 0)_2$$

and in quaternary format as:

$$(0331122021302)_Q$$

Actually this is a discrete signal which represents the given DNA sequence. Figure 1 explains the graphical representation of the above sequence.

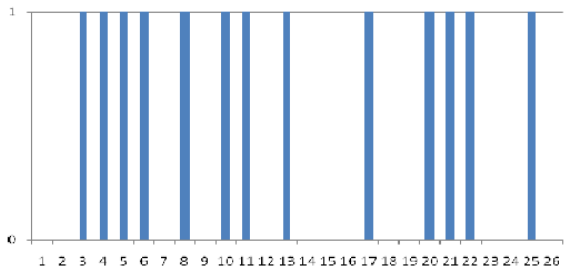

Fig.1: Barcode representation of the DNA sequence.

As observed from the this figure, the binary representation of the DNA sequence is nothing but a barcode. Therefore, we can use a bar code reader to read this code. The search of the DNA sequences will be vastly accelerated if we use this binary representation in the database instead of string matching. However, barcodes will be quite long in length for most DNA strands [4, 5].

3. DNA Matrix and Polynomial Representations

In section 3.1 we will represent the DNA sequence as a two-dimensional signal image. In other words, we will use matrix representation. In section 3.2, we will discuss a sparse polynomial DNA representation. In sections 3.3 and 3.4, we will discuss the formulation as quick response format.

3.1 Matrix Representation

The matrix representation of a DNA sequence consists of four rows. Each row represents the occurrence of one DNA base (A, C, G or T).

For example the DNA sequence with length 13 base:

ATTCCGGAGCTAG

It will become in matrix representation as:

$$\begin{array}{c} A \\ C \\ G \\ T \end{array} \begin{pmatrix} 1000000100010 \\ 0001100001000 \\ 0000011010001 \\ 0110000000100 \end{pmatrix}$$

As we examine the above shown matrix, the first row represents the occurrences of the (A) molecule in the DNA sequence and the second, third and fourth rows represent the (C), (G) and (T) occurrences respectively.
We can visually represent this matrix as a dot for each occurrence of the digit one and a blank for each occurrence of the digit zero.

3.2 Formulation as sparse polynomial

The DNA matrix representation is actually a sparse matrix. The matrix contains four sparse vectors for each base (A, C, G and T). We can utilize these row vectors to generate four sparse polynomials as follows:

$$P(x) = \sum_{i=0}^{n} a_i x^i$$

Where a_i is ε {0, 1} and n is the Genome length. For example the previously discussed matrix will result in the following sparse polynomials:

$P_A = X^0 + X^7 + X^{11}$
$P_C = X^3 + X^4 + X^9$
$P_G = X^5 + X^6 + X^8 + X^{12}$
$P_T = X^1 + X^2 + X^{10}$

Consequently, we will be able to store the power of each term in each polynomial instead of the whole sequence as follows:

$P_A = (0, 7, 11)$
$P_C = (3, 4, 9)$
$P_G = (5, 6, 8, 12)$
$P_T = (1, 2, 10)$

The length of the DNA sequence may reach millions of bases. For example the human genome weight is 3.2pg where each pg is equivalent to 921 million bases. Therefore, the human genome size (in bp) = (0.921×10^9) x DNA content (pg) = (0.921×10^9) x 3.0 (pg) = 2763 Mb. By no means, the human genome is the largest in weight. For example, a Japanese plant was recently discovered to have a weight of 132.5pg. In other words, this plant has about 44 times the human genome base pairs. The resulting sparse polynomials powers will be constituted of very large numbers. Storing the difference between the powers appreciably reduces the required number of storage bits. This is a well known technique in simple lossless data compression. For example, and referring to the previously discussed DNA string, we get

$P_{dA} = (0, 7, 4)$
$P_{dC} = (3, 1, 5)$
$P_{dG} = (5, 1, 2, 4)$
$P_{dT} = (1, 1, 8)$

These four polynomials can be augmented in one polynomial by adding multiples of the DNA string length to the elements and the result will be

P = (0, 7, 11, 16, 17, 22, 31, 32, 34, 38, 40, 41, 49)

The final difference sparse polynomial is given by:

$P_d = (0, 7, 4, 5, 1, 5, 9, 1, 2, 4, 2, 1, 8)$

The elements of this last vector can be stored in 52 bits compared to the original 104 bits with a resulting compression ratio of 1/2. Using sparse polynomial representation for DNA sequences speeds up the searching time in the database; each sequence in the database will be stored as four sparse polynomials (P_{dA}, P_{dC}, P_{dG} and P_{dT}), the query sequence will be presented in four sparse polynomials too (P_{dqA}, P_{dqC}, P_{dqG} and P_{dqT}) and finally we find the hamming distance between each adjacent vector (h_A ,h_C ,h_G and h_T).

$$H = h_A + h_C + h_G + h_T$$

The vector with the minimum total hamming distance (H) will be the nearest one to the query sequence.

3.3 Formulation as a Quick Response Format

A QR code is a visual matrix code or a two-dimensional code that is readable by dedicated QR barcode readers and phone cameras. The code consists of black modules arranged in a square pattern on a white background. The information encoded can be text, URL or other data. Figure 2 shows a sample commercial QR code image.

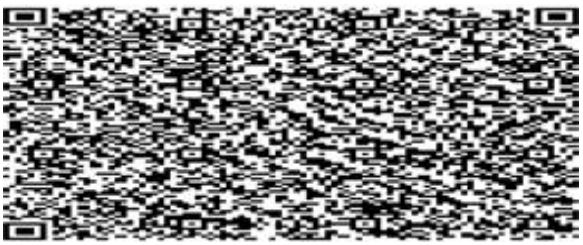

Fig.2: QR code

These QR are more useful than a standard barcode. They can store much more data and are

very compact in size. The other key feature of QR Codes is that instead of requiring a chunky laser hand-held scanner, many modern mobile phones can be adapted to scan them. We can store the QR image of the DNA sequence instead of the sequence itself in the database, therefore it will be more compact and the search performance will be much faster than if we use other image processing technique [6].

3.4 Barcode versus QR code

There are many advantages to use QR code when compared to standard barcodes. The QR code storage capability vastly exceeds that of the Barcode. Moreover, the QR code is a two dimensional code not a single dimensional as the barcode, therefore the reading process in case of QR code will be performed in a parallel way not sequentially as the barcode reader. Finally, the QR code reader can be implemented using mobile phone cameras and there is no need for expensive barcode readers that may not be portable, and finally the QR code reader is actually an image reader and can also read the barcode.

4. Methodology

In the previous section, we have represented the DNA sequence as a single dimension binary signal. Hence, we can apply the Haar wavelet on the DNA signal. For example the DNA sequence

ATTCCGTA

which is in binary format:

$$(0011110101101100)_2$$

and after applying the first round of Haar wavelet it will become

$$(0,\sqrt{2},\sqrt{2},1/\sqrt{2},\ 1/\sqrt{2},\ 1/\sqrt{2},\sqrt{2},0)$$

The second round of the Haar wavelet on the trend signal only it will be give

$$(1,\ 2,\ \sqrt{2},\ 1)$$

By applying the Haar wavelet until the trend will be one term only; all the previous signals will be stored in the DNA database instead of the sequence itself. In the same way we apply the Haar wavelet on the query sequence and the search process will be done on the trend signals instead of the whole genome.

4.1 Haar Wavelet of the DNA Signal

As was briefly discussed, this method divides the database into clusters with the same trends. These clusters can be regarded as classification groups that enable the researcher to identify the group of the query sequence. This is a major advantage of the proposed methodology.

It is a DNA mining technique. The search process will go in a bottom-up fashion. It starts to compare the smallest trend signals and if it matches, the search goes upward and this process is repeated until we get the exact match. The database is acting similar to a large tree and the search process will be like searching through this tree branches. Figure 3 presents the search process in the database starting from the tree root or largest trend.

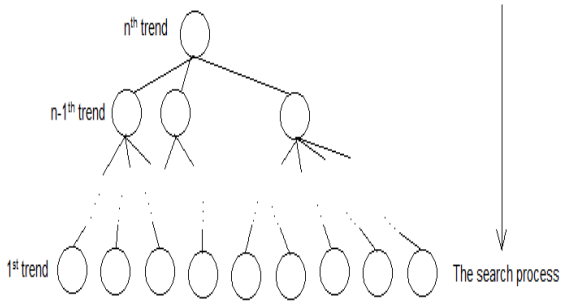

Fig.3: The search process in the DNA database.

The previous tree is actually a B-tree; therefore searching a B-tree is O ($t\log_t$ (n)) where **t** is the block size. As observed O ($t\log_t$ (n)) < O (n); therefore this method provides faster search times as compared to the traditional string matching methods [10].

4.2 Verification

The recent publication of the complete chimp genome [11], marked by a celebratory issue of

the journal "Nature" recounts us that humans and chimps share 96 percent of the same genetic material. For example, consider the first 100 base pairs of the chimps mitochondrial DNA and for humans.

Chimps:
GTTTATGTAGCTTACCCCCTCAAAGCAATACACTGAAAAT
GTTTCGACGGGTTTACATCACCCCATAAACAAACAGGTTT
GGTCCTAGCCTTTCTATTAG

Humans:
GATCACAGGTCTATCACCCTATTAACCACTCACGGGAGCT
CTCCATGCATTTGGTATTTTCGTCTGGGGGGTGTGCACGC
GATAGCATTGCGAGACGCTG

Applying Haar wavelets on the two strands we will find that the results of the last trend are very close.

Chimps:
$trend_{n-2}$ = (3.27036886298778, 2.3864853865046, 2.56326208180123, 0.61871843353823)
$trend_{n-1}$ = (3.99999999999999961967979879025, 2.24999999999997270706559232398 3)
$trend_n$ = (4.41941738241592)

Humans:
$trend_{n-2}$= (2.56326208180123, 2.38648538650459, 3.62392225358105, 0.265165042944955)
$trend_{n-1}$= (3.49999999999991040387217562 3028, 2.74999999999995485685603750 9077)
$trend_n$ = (4.4194173824159125)

As observed from the previous example, the two strands belong to the same evolutionary group ($trend_n$). To get more accurate results one has to go back one step then identify the sequence if it whether is Chimp or Human. In this work, we have developed a C# program, as shown in Appendix A, which transforms the DNA sequence to QR code. The program first transforms the DNA sequence to its matrix representation and then plots a dot for each 1 and blank for each 0. We apply the program on a bacterium sequence of length 32767 bases and the result image was quite compact. Figure 4 demonstrates the QR of the selected DNA sequence.

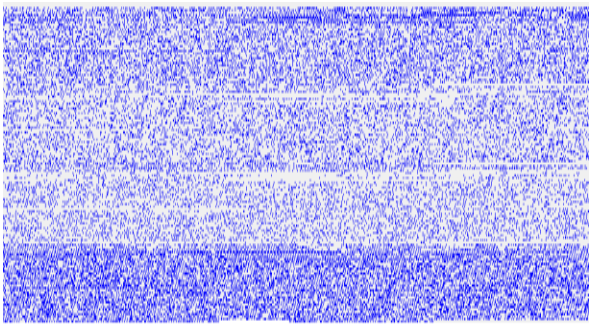

Fig.4: The QR code for a DNA sequence of length equal to 32767 bases.

The QR code is a compact image that lends itself to image processing techniques. Applying the Haar wavelet on the same example of the Chimps and Humans, Figure 5 explains that the Chimps and Humans belong to the same evolutionary family. The fourth trend can be thought of as the species family and the various levels as the sub-species and so on until reaching the individual using the original DNA signal. As a matter of fact, this type of classification is the numeric and visual representations of taxonomic species groups. Studying the visual representation of the DNA provides an insight into the evolutionary process of a specific species.

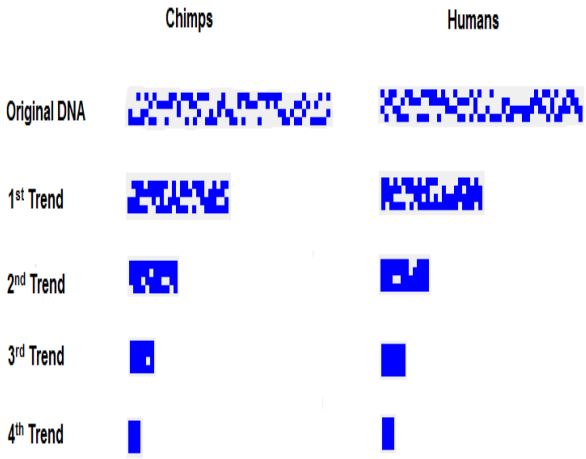

Fig.5: The QR code comparisons for DNA sequences of Chimps and Humans.

As observed from the previous figure that the fourth trend signal representations of the QR codes of Chimps and Humans are identical. This demonstrates the previous notions of evolutionary groups.

Summary and conclusion

Based on the previous discussion, one establishes the following:

- Storing the DNA sequences in database as QR code images has more practical advantages than storing the sequences themselves.
- Applying Haar wavelet on the visual representation of the signals and searching in the Haar trend signals is much faster than searching the DNA sequences themselves.
- The search process is built on image comparisons rather than base matching or what is known as string matching.
- Based on the Haar transform, the method provides a numeric value that identifies evolutionary groups.
- Based on the discussions of sections 3.2 and 4.1, a considerable improvement in search speeds and the minimization of storage requirements are both achievable by utilizing the proposed technique.

References

[1] M. Saeb, E. El-Abd, M. E. El-Zanaty, "DNA Steganography using DNA Recombinant and DNA Mutagenesis Techniques," WSEAS Transaction on Computer Research, Vol. 2, pp. 50 -56, 2007.

[2] Wendell M. Smith, "DNA Based Steganography for Security Marking," Xix International Security Printers' Conferences, Montreux, 14-16 May, 2003.

[3] Ashish Gehani, Thomas LaBean, and John Reif, "DNA-Based Cryptography," IMACS DNA-Based Computers V, American Mathematical Society, US., 2000.

[4] Site: barcodeoflife.org "what is DNA Barcoding,http://www.barcodeoflife.org/content/about/what-dna-barcoding, 2010.

[5] Rasmus Nielsen Mikhail Matz, "Statistical Approaches for DNA Barcoding," Society of Systematic Biologists', US., 2006.

[6] Ching-yin Law and Simon So, "QR codes in Education," Journal of Educational Technology Development and Exchange', Volume 3, No. 1, pp. 85-100, Hong Kong, October 2010.

[7] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ, "Basic local alignment search tool". J Mol Biol 215 (3): 403–410, October 1990 instead of wikipedia

[8] Adrian D. Haimovich, Bruce Byrne, Ramakrishna Ramaswamy, William J. Welsh, "Wavelet Analysis of DNA Walks," Journal of Computational biology, Vol. 13,Number 7, pp. 1289-1298, 2006.

[9] A. Arneodo, Y. D'Aubenton-Carafa, B. Audit, E. Bacry, J.F. Muzy, C. Thermes, "What can we learn with wavelets about DNA sequences?," Elsevier Science, Physics A 249, pp. 439-448, 1998.

[10] A. Arneodo, Y. D'Aubenton-Carafa, B. Audit, E. Bacry, C Vaillant, C. Thermes, " Extracting structural and dynamical information from wavelet-based analysis of DNA sequences," Coll. Group Theoretical Methods in Physics, IOP Publishing Ltd, France, 2003.

[11] Eamonn Keogh, Stefano Lonardi, Victor B. Zordan, Sang-Hee Lee, Manel Jara, "Visualizing the Similarity of Human and Chimp DNA," University of California, Riverside, USA, 2005.

Appendix A: DNA as QR code

In the following pseudo code, we describe the proposed procedure in a formal approach.

```
Algorithm of drawing the QR code of DNA
[Given a DNA sequence D-M [4] [D-length]]
Input: D;
Output: QR code image;
Algorithm Body:
For i = 1 to D-length
// loop for all bases of the DNA sequence.
        Switch Case D[i]:
                Case 'A'
                        M[0][i] := 1;
                Case 'C'
                        M[1][i] := 1;
                Case 'G'
                        M[2][i] := 1;
                Case 'T'
                        M[3][i] := 1;
Next i
image QR := Draw(M);
// Function to draw the two dimensional image
matrix
Draw(Matrix[][]);
 Begin function
    for  i = 1 to Matrix-length
        for  j = 1 to Matrix [i].length
            if (Matrix[i][j] == 1)
                DrawCircle(blue,i,j,1,1);
        next j;
     next i;
 End function
```