DNA Sequence Representation and Comparison Based on Quaternion Number System

Hsuan-T. Chang, Chung J. Kuo
Photonics and Information Lab,
Department of Electrical Engineering, YunTech
Douliu Yunlin, 64002 Taiwan

Neng-Wen Lo
Department of Animal Science and Biotechnology,
Tunghai University
Taichung, 40704 Taiwan

Wei-Z.Lv
Computer and Communication Lab,
Industrial Technology Research Institute
Hsinchu, 310 Taiwan

***Abstract*—Conventional schemes for DNA sequence representation, storage, and processing areusually developed based on the character-based formats.We propose the quaternion number system for numerical representation and further processing on DNA sequences.In the proposed method, the quaternion cross-correlation operation can be used to obtain both the global and local matching/mismatching information between two DNA sequences from the depicted one-dimensional curve and two-dimensional pattern, respectively.Simulation results on various DNA sequences and the comparison result with the well-known BLAST method are obtained to verify the effectiveness of the proposed method.**

Keywords- Bioinformatics; genomic signal processing; DNA sequence; quaternion number; data visualization.

I. INTRODUCTION

Recently, the great progress on biotechnology makes the deoxyribonucleic acid (DNA) sequencing more efficiently. Huge amount of DNA sequences of various organisms have been successfully sequenced with higher accuracy. Analyzing DNA sequences can investigate the biological relationships such as homologous andphylogeny of different species. However, the analysis of DNA sequences using the biological methods is too slow for processing huge amount of DNA sequences. Therefore, the assistance of computers is necessary and thus *bioinformatics* is extensively developed. Efficient algorithms are desired to deal with the considerable and tedious biomolecular data.

Computer-based algorithms have solved various problems dealt in bioinformatics, such as the sequence matching (two and multiple sequences, global and local alignments), fragments assembly of DNA pieces, and physical mapping of DNA sequences. Most of the algorithms consider the data structures of DNA sequences as the string, tree, and graph. The artificial intelligence techniques such as the genetic algorithm [1], artificial neural networks [2], and data mining [3] have been intensively employed in this research area. In [4], the study of genomic signals mainly at scales of 104~108 bp, to detect general trends of the genomic signals, potentially significant in revealing their basic properties and to search for specific genomic signals with possible control functions. On the other hand, many distributed databases over the Internet have been constructed and can be easily accessed from the World Wide Web [5]-[7].Most of the techniques treat the DNA sequences as the symbolic data, which are the composition of four characters A, G, C, and T corresponding to the four types of nucleic acids: Adenine, Guanine, Cytosine, and Thymine, respectively. However, the biomolecular structures of genomic sequences can be represented as not only the symbolic data but also the numeric form.

Recently the genomic signal processing era has received a great deal of attention [8], [9].The well-known digital signal processing (DSP) techniques have been developed to analyze the numeric signals for many applications [10].If the symbolic DNA sequences can be transformed to the numeric ones, then the DSP-based algorithms would provide alternative solutions for the bioinformatics problems defined in the symbolic domain. Hsieh *et. al.* proposed DNA-based schemes to efficiently solve the graph isomorphism problems [11], [12]. Some previous studies have shown various methods of mapping the symbolic DNA sequences to numeric ones for further processing such as discrete Fourier transform or wavelet transform [13]-[15].Then, the periodic patterns existed in DNA sequences can be observed from the determined scalograms or spectrograms. In assigning the four bases as some real or complex numbers such as $\pm 1 \pm i$, the further mathematical operations are straightforward and simple.However, exploring the biological relationship is difficult in the mapping between the bases and numbers.

Magarshak proposed a quaternion representation of RNA sequences [16].Four bases with eight biological states form the group of quaternions and the tertiary structure of the RNA sequence can be analyzed through the quaternion formalism. Hypercomplex signals can be considered as the general form of quaternion signals [17]. Shu*et. al.* proposed hypercomplex number representation for pairwise alignment and determining the cross-correlation of DNA sequences [18]-[20].The DNA sequences are aligned with fuzzy composition and a new scoring system was proposed to adapt the hypercomplex number representation. Based on the similar ideas shown above, we adapt two implications using the quaternion number system [21] for DNA sequences. The quaternion numbers are complex numbers with one real part and three imaginary parts. Here four bases in a DNA sequence are assigned with four different quaternion numbers. A DNA sequence can thus be transformed into a quaternion-number sequence. Instead of finding the local frequency information of DNA sequences, the cross-correlation algorithm based on quaternion numbers is proposed for both the global and local matching between two DNA sequences. Since the real parts of four quaternion

numbers are all zero and the imaginary parts are with certain properties. The matching information can be observed from the real part in the result of the quaternion cross-correlation operation. In addition to the global cross-correlation result, the local matching information can be extracted from the product of each multiplication in the correlation operation. The global and local comparisons can be represented by 1-D curve and 2-D pattern, respectively. The simulation results show that the proposed quaternion number system can efficiently represent DNA sequences and then be used to determine the global and local sequence alignment with the help of the cross-correlation operation.

The organization of this paper is as follows: Section 2 introduces the basics of quaternion number systems and the quaternion number representation of DNA sequences. The global and local sequence matching based on the quaternion correlation is described. Section 3 provides the simulation results for certain DNA sequences, which verify the effectiveness of the proposed method. Finally, the conclusion is drawn in Section 4.

II. Quaternion Correlation For Sequence Matching

A. Quaternion Number Sequence Representation

Quaternion numbers [21] (also called the *hyper-complex numbers*) are the generalization of complex numbers. They have been applied in certain applications such as the color image filtering [22] and segmentation [23] and, the design of 3-D infinite impulse response filters [24].Since the quaternion number system is not well known in all signal processing areas, here their properties are briefly reviewed.

The quaternion number has four components: one real part and three imaginary parts. The notation of a quaternion number q is defined as

$$q = q_a + \hat{i}q_b + \hat{j}q_c + \hat{k}q_d \tag{1}$$

Where q_a, q_b, q_c, q_d are real numbers, and $\hat{i}, \hat{j}, \hat{k}$ are operators for the three imaginary parts. The conjugate of quaternion number, q^*, is defined as

$$q^* = q_a - \hat{i}q_b - \hat{j}q_c - \hat{k}q_d . \tag{2}$$

More detailed description of quaternion operations can be referred in [25].

Table I shows the proposed mapping method between the four characters and the corresponding quaternion numbers. Note that all the real parts are zero, while the imaginary parts can be considered as the coordinates of four vertices of a regular tetrahedron in 3-D space.

That is, the 3-D coordinates (x,y,z) of the vertices of a regular tetrahedron are applied to the imaginary part `$b$,$c$,$d$' of quaternion numbers, and the real part `a' of quaternion numbers are set to zero. Then, a character sequence is mapped to a quaternion number sequence. For example, a character sequence $s_c[n]$={A,A,T,A,G,C,G,T} is mapped to a quaternion number sequence $s_q[n]=\{q_A, q_A, q_T, q_A, q_G, q_C, q_G, q_T\}$.

TABLE I: THE MAPPING TABLE FROM A, C, T, AND G TO CORRESPONDING QUATERNION NUMBERS.

	q_a	q_b	q_c	q_d
A	0	0	0	1
T	0	$\frac{2\sqrt{2}}{3}$	0	$-\frac{1}{3}$
C	0	$-\frac{\sqrt{2}}{3}$	$\frac{\sqrt{6}}{3}$	$-\frac{1}{3}$
G	0	$-\frac{\sqrt{2}}{3}$	$-\frac{\sqrt{6}}{3}$	$-\frac{1}{3}$

After the mapping procedure in the former discussion, a quaternion number sequence standing for a DNA sequence is obtained:

$$s_q[n] = \{q_1, q_2, q_3, ...,q_n\}. \tag{3}$$

Then, an accumulating process is applied to obtain another quaternion number series

$$\bar{s}_q[n] = \{\bar{q}_1, \bar{q}_2, \bar{q}_3, \mathrm{K}\ , \bar{q}_n\}, \tag{4}$$

where

$$\bar{q}_l = \sum_{n=1}^{l} q_n . \tag{5}$$

By extracting the imaginary part of series $\bar{s}_q[n]$, the 3-D coordinates for each accumulated quaternion number in the series can be obtained. Line segments are used to connect these points in order, and a 3-D trajectory can thus be obtained for DNA sequence visualization [26]-[28].

B. Quaternion Correlation

Once the quaternion number sequences of two DNA sequences have been obtained, the cross-correlation operation is performed for the global comparison. The cross-correlation of two quaternion number sequences $s_{q_1}[n]$ and $s_{q_2}[n]$ of lengths M and N, respectively, is defined as

$$r_{1,2}[\xi] = \sum_{n=-\infty}^{\infty} s_{q_1}[n] s^*_{q_2}[n+\xi], \tag{6}$$

where ξ is the index of the correlation function and $0 \le \xi \le N + M - 1$.

In the correlation operation, the conjugate operation is applied on one of the two quaternion numbers multiplied. Therefore, the cross-correlation of two identical and different symbols contributes +1 and -1/3 to the result, respectively. If there are two sequences to be correlated with length N and M (N>M), the real part of correlation result, v_{Re}, forzbp overlap (0<$z \le M$) is

$$v_{\mathrm{Re}} = p + (z - p) \times (-1/3), \tag{7}$$

Where p is the matching counts and z-p is the mismatching counts.

Equation (6) shows that the cross-correlation result for a specific ξ value is the sum of the product of the original and the shifted and conjugate sequences. The products in certain n are non-zero and zero values when two sequences overlap and not, respectively.

When the two sequences overlap, the product of two quaternion numbers reflects whether they are the same or not. That is, during the cross-correlation operation, the local alignment proceeds under a given ξ value. For example, if there are z overlapping numbers between two sequences, Eq. (6) becomes

$$r_{1,2}[\xi] = \sum_{n=m}^{m+z-1} s_{q_1}[n] s_{q_2}^{*}[n+\xi], \text{ for } z \geq 1, \tag{8}$$

Where m denotes the starting position of the overlap region under the given ξ value. Since there are (N+M-1) possible ξ values in the cross-correlation result, all the possible local alignments between two sequences can be obtained. Therefore, the local alignment results can be shown in a 2-D array of size (N+M-1)×(N+M-1), in which the entry denotes the matching status of two nucleotides from two DNA sequences. A grayscale image $f_{\mathrm{Re}}(x,y)$ of size (N+M-1) × (N+M-1) corresponding this 2-D array can be generated based on the following rule:

$$f_{\mathrm{Re}}(x,y) = \begin{cases} 0, & \text{if } \mathrm{Re}\{s_{q_1}[n] s_{q_2}^{*}[x+y]\} = 1, \quad \text{(overlap \& match)} \\ 255, & \text{if } \mathrm{Re}\{s_{q_1}[n] s_{q_2}^{*}[x+y]\} = -\frac{1}{3}, \quad \text{(overlap but mismatch)} \\ 128, & \text{if } \mathrm{Re}\{s_{q_1}[n] s_{q_2}^{*}[x+y]\} = 0, \quad \text{(non - overlap)} \end{cases} \tag{9}$$

Here Re{·} denotes the real part of the complex value in the bracket and 0<x, y≤N+M-1. If the connected pixels in the horizontal direction constitutes as a black line in the image, the local matching between two sequences exists. Therefore, the local matching information between two DNA sequences can be obtained from the quaternion correlation result in addition to the global matching information.

TABLE II. THE IMAGINARY PARTS OF THE MULTIPLIED VALUES OF EVERY TWO DIFFERENT QUATERNION NUMBERS.

	AxT	AxC	AxG	TxC	TxG	CxG
$\hat{i}$	0	$-\frac{\sqrt{6}}{3}$	$\frac{\sqrt{6}}{3}$	$\frac{\sqrt{6}}{9}$	$-\frac{\sqrt{6}}{9}$	$-\frac{2\sqrt{6}}{9}$
$\hat{j}$	$\frac{2\sqrt{2}}{3}$	$-\frac{\sqrt{2}}{3}$	$-\frac{\sqrt{2}}{3}$	$\frac{\sqrt{2}}{3}$	$\frac{\sqrt{2}}{3}$	0
$\hat{k}$	0	0	0	$\frac{4\sqrt{3}}{9}$	$-\frac{4\sqrt{3}}{9}$	$\frac{4\sqrt{3}}{9}$

	TxA	CxA	GxA	CxT	GxT	GxC
$\hat{i}$	0	$\frac{\sqrt{6}}{3}$	$-\frac{\sqrt{6}}{3}$	$-\frac{\sqrt{6}}{9}$	$\frac{\sqrt{6}}{9}$	$\frac{2\sqrt{6}}{9}$
$\hat{j}$	$-\frac{2\sqrt{2}}{3}$	$\frac{\sqrt{2}}{3}$	$\frac{\sqrt{2}}{3}$	$-\frac{\sqrt{2}}{3}$	$-\frac{\sqrt{2}}{3}$	0
$\hat{k}$	0	0	0	$-\frac{4\sqrt{3}}{9}$	$\frac{4\sqrt{3}}{9}$	$-\frac{4\sqrt{3}}{9}$

C. Mismatching Analysis

The real part of correlation result reflects the matching information, while the imaginary part reflects the mismatching information in sequence comparison. Therefore, the number of the matching and mismatching counts from the real and imaginary parts of the correlation result could be investigated. Let q_3 and $q_4$denote the products of two quaternion numbers q_1 and q_2. That is, q_3=q_1 × q_2 and q_4=q_2 × q_1,where $q_n = q_{a_n} + \hat{i} q_{b_n} + \hat{j} q_{c_n} + \hat{k} q_{d_n}$, n=1,2,3,4. According to the rules of quaternion multiplication, the following results can be obtained:

$$q_{a_3} = -q_{b_1} \times q_{b_2} - q_{c_1} \times q_{c_2} - q_{d_1} \times q_{d_2}, \tag{10}$$

$$q_{b_3} = q_{c_1} \times q_{d_2} - q_{d_1} \times q_{c_2}, \tag{11}$$

$$q_{c_3} = -q_{b_1} \times q_{d_2} + q_{d_1} \times q_{b_2}, \tag{12}$$

$$q_{d_3} = q_{b_1} \times q_{c_2} - q_{c_1} \times q_{b_2}, \tag{13}$$

and

$$q_{a_4} = -q_{a_3},\ q_{b_4} = -q_{b_3}, q_{c_4} = -q_{c_1}, q_{d_4} = -q_{d_3}. \tag{14}$$

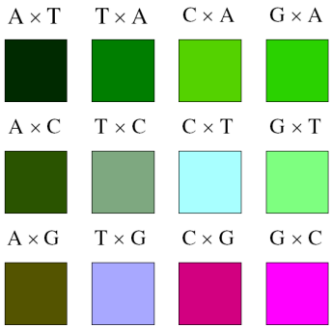

Figure 1.The colors corresponding to the different combination of the multiplied quaternion numbers.

In addition to the real part of the calculated quaternion number, the rest imaginary parts can provide further more mismatching information between the two sequences. A color image in which the R, G, and B components are respectively corresponding to the three imaginary parts $\hat{i}, \hat{j}, \hat{k}$ in the quaternion number can be generated.

First, the possible values of the three imaginary parts are determined and listed in Table I. Second, for each component, the finite values are normalized into the monotone pixel values between 0 and 255.

There are six, five, and three possible values for the $\hat{i}$, $\hat{j}$, and $\hat{k}$ imaginary parts, respectively. Each value in each imaginary part can be assigned to as a grayscale value. Therefore, three grayscale images for the R, G, and B components can be generated. By combining the three components, different colors corresponding to the various mismatching conditions between two nucleotides can be obtained. Figure 1 shows the different colors and the corresponding mismatching conditions. According to the color assignments in Fig. 1, a 2-D pattern $f_{\mathrm{Im}}(x,y)$ that reflects the imaginary parts of the correlation results can be generated. The image $f_{\mathrm{Im}}(x,y)$ is similar to the image $f_{\mathrm{Re}}(x,y)$ that reflects

the real parts of the correlation results. However, each pixel represents one of the different mismatching conditions.

D. Complexity Analysis

General performance measurements of an algorithm are the time/computation and space complexity. By definition, the correlation of two sequences $x_1[n]$ and $x_2[n]$ of size N needs N^2 multiplication and N-1 addition if the two sequences are of length N. The time complexity of the original correlation is $O(N^2)$. In DSP theories, the discrete Fourier transform (FT) is used to accelerate the speed of correlation (the correlation theorem).That is,

$$r_{1,2}[\xi] = \sum_{l=0}^{n-1} x_1[n] x_2^* [l+\xi] = \mathrm{FT}^{-1}\{X_1^*[\kappa] X_2[\kappa]\}, \quad (15)$$

where $X_1[\kappa] = \mathrm{FT}\{x_1[n]\}$ and $X_2[\kappa] = \mathrm{FT}\{x_2[n]\}$ are the FTs of two sequences $x_1[n]$ and $x_2[n]$, respectively. In Eq. (15), the time complexity depends on the FT. Because the time complexity of the FT is $O(N^2)$ and time complexity of multiplication is $O(N)$, the time complexity of correlation operation is $O(N^2)$. With the fast FT algorithm, the time complexity can be improved to become $O(N\log_2 N)$.

Let FT_Q and $\mathrm{FT}_\mathrm{Q}^{-1}$ denote the quaternion Fourier transform (QFT) and inverse QFT, respectively. From the Hypercomplex Wienear-Khintchine theorem [23], Eq.(15) becomes:

$$r_{1,2}[\xi] = \mathrm{FT}_\mathrm{Q}^{-1}\{X_1[\kappa] X_{2\|}^*[\kappa]\} + \mathrm{FT}_\mathrm{Q}\{X_1[\kappa] X_{2\perp}^*[\kappa]\}, \quad (16)$$

where $X_1[\kappa] = \mathrm{FT}_\mathrm{Q}\{x_1[n]\}$, $X_2[\kappa] = \mathrm{FT}_\mathrm{Q}\{x_2[n]\}$, $X_{2\|}[\kappa] \| \hat{\mu}$, and $X_{2\perp}[\kappa] \perp \hat{\mu}$ .Note that $\hat{\mu}$ is the unit pure quaternion and it is referred to as the eigen axis, which represents the direction in the 3-D space of imaginary part of a quaternion. A QFT can be implemented by two ordinary FTs [29]. That is,

$$r_{1,2}[\xi] = \mathrm{FT}^{-1}\{X_{1v}[\kappa] X_{2u}^*[\kappa] - X_{1u}[-\kappa] X_{2v}[-\kappa]\}, \quad (17)$$

Where $x_u[n]= x_a[n]+ix_b[n]$,$x_v[n]= x_c[x]-ix_d[n]$.Therefore, the time complexity of quaternion correlation can be significantly reduced to$O(N\log_2 N)$.

Regarding to space complexity, the most memory-consumable is the storage of numerical data for DNA sequences. If the correlation is calculated by using the QFT, it needs additional memory space to store data of frequency domain in the computational process. Therefore, this method trades the memory space for efficiency.

III. Experimental Results

In computer simulation, computer-generated random sequences and real DNA sequences are used to perform the quaternion correlation. Consider a random quaternion sequence $s_{x1}[n]$ and let $s_{x2}[n]$ denote the prefix (first eight quaternion numbers) of the sequence $s_{x1}[n]$.Following the cross-correlation operation, the cross-correlation result is also a quaternion number sequence. Figure 2 shows the real part of the correlation results, which distribute over the eight discrete levels. Actually, there are nine situations in the correlation results from the matching counts being zero to eight. By observing the coefficients, it shows that each level differs by 4/3.Therefore, the real parts of cross-correlation coefficients are relative to the matching counts between two sequences. There is a maximum correlation at the position zero because this is the exactly matching position.

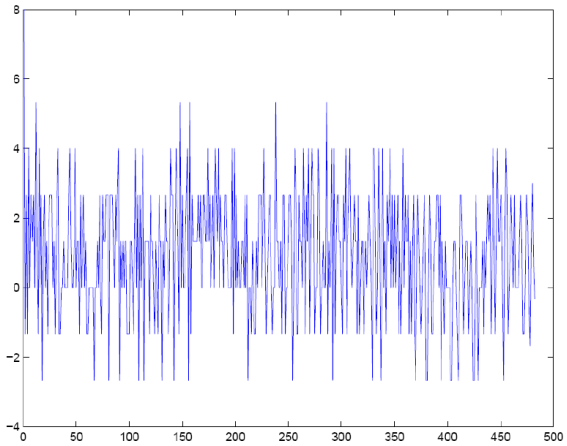

Figure 2. The real-part correlation coefficients of two quaternion sequences $s_{x1}[n]$ and let $s_{x2}[n]$.

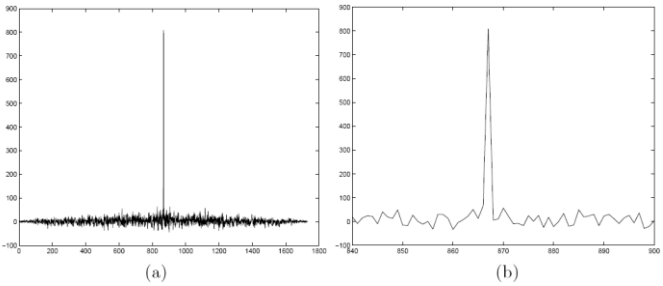

Figure 3. The real-part correlation coefficients of the sequences for the TGFA genes of Human and Mouse. (a) The highest correlation peak appears at the best-matching position; (b) The detailed center region.

Three DNA sequences retrieved from the web-based databases of National Center for Biotechnology Information (NCBI) [30] are then used to test the proposed method. Each sequence has an accession number for identification. First of all, consider two sequences from highly similar genes: the human TGFA sequence (Accession: K03222, 867 bp) and the mouse TGFA sequence (Accession: BC003895, 1024 bp).The DiHydro Folate Reductase (DHFR) gene (Accession: L26316, 1042 bp) from a mouse is also considered. Figure 3(a) shows the quaternion cross-correlation result of the two TGFA genes. A correlation peak with value 807 appearing at the position ξ =867 represents that there exists large similarity (822 identical base pairs) and the best matching position of two sequences can be obtained. Figure 3(b) shows a detailed region from ξ =840 to 900.The correlation values at other positions are much smaller than the peak value. Therefore, it is verified that the proposed method can determine the best global matching position of two sequences. On the other hand, consider the cases of base-pair deletion and insertion, which commonly happens in DNA sequences. To investigate the effects, the sequence of mouse TGFA gene is modified and then used to determine the correlation result. Figure 4(a) and 4(c) show the

correlation results when 1-bp deletion and insertion happen in the center position of the TGFA gene, respectively. As shown in Figs. 4(b) and 4(d) for details, two half values of the original correlation peak value appear at the deletion/insertion positions.

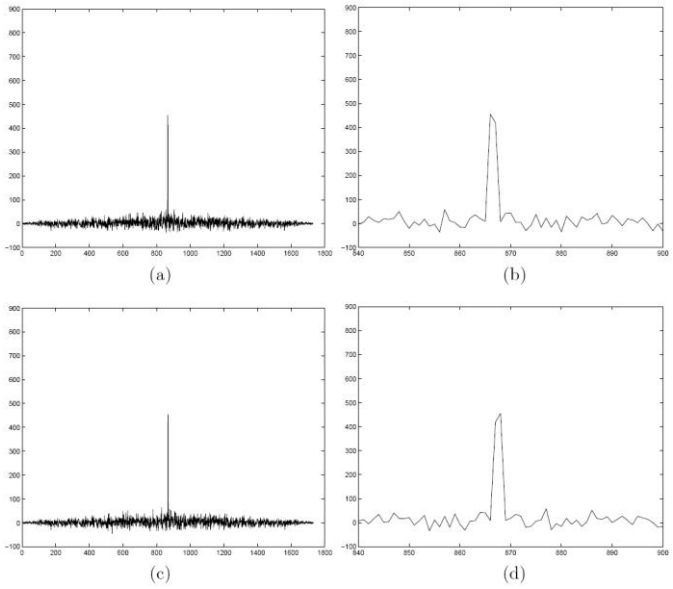

Figure 4. The correlation results for (a) 1-bp deletion in the center position of the Human TGFA sequence; (b) the detailed center region in (a); (c) 1-bp insertion in the center position of the Human TGFA sequence;(d) the detailed center region in (c).

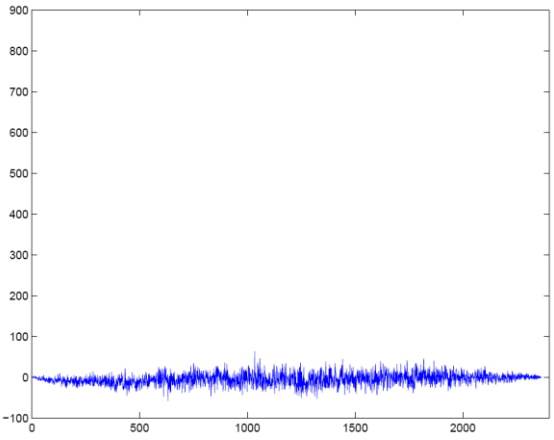

Figure 5. The real-part of cross-correlation coefficients of the sequences for the Human TGFA and Mouse DHFR genes.}

According to the peak value in the correlation result, the number of matched base pairs can be estimated by the use of Eq.(7).If the number of insertion/deletion increases, the corresponding correlation peak values decrease. The partial matching positions can still be detected from the decreased correlation peaks. Therefore, the deletion or insertion in sequences can be estimated from the correlation result. Figure 5 shows the correlation result of two quite different (human TGFA and mouse DHFR) genes. There is no significant peak value among all the real-part coefficients. It verifies that the similarity between these two gene sequences is small.

The sequence matching results obtained from quaternion correlation are compared with the well-known BLAST method [31].Figure 6 shows the top 34 sequences retrieved by the use of BLAST when using the Human TGFA gene (Accession: K03222, 867 bp) as the query sequence. The sequences for the first 10 high scores (excluding the query sequence itself) are used to perform the quaternion cross-correlation with the query sequence. Each cross-correlation result shows a correlation peak at the best-matching position. Table III summarizes the correlation results and BLAST scores for comparison. The peak values almost follow the trend of BLAST scores. Since BLAST is designed specifically for local alignment of sequences, the small difference between the scores and peak values is reasonable.

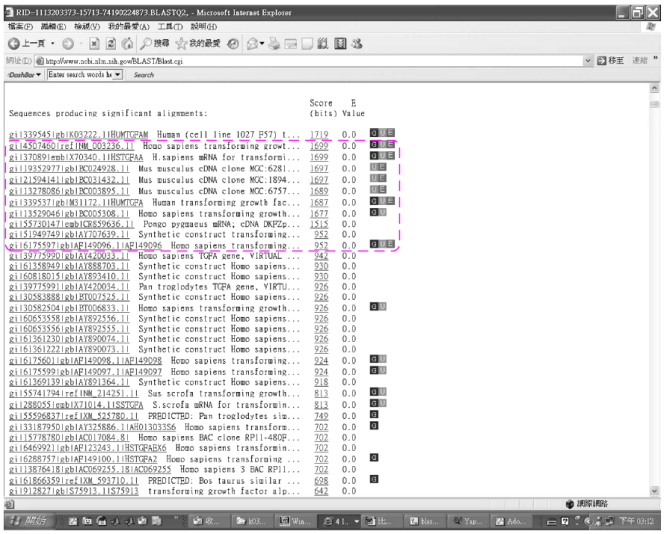

Figure 6. The query result of the Human TGFA gene using the Nucleotide-nucleotide BLAST (blastn) tool.}

TABLE III. THE CORRELATION PEAK VALUES AND POSITIONS AND CORRESPONDING TOP TEN SCORES WHEN USING THE BLASTN TO QUERY THE HUMAN TGFA SEQUENCE IN NCBI DATABASE.

Sequence information	Position	Peak value	Blast score
Homo sapiens TGFA, K03222, 867 bp	867	867	1719
Homo sapiens TGFA, NM_003236, 4119 bp	4122	860	1699
Homo sapiens mRNA for TGFA, X70340, 4119 bp	4122	860	1699
Mus musculus cDNA clone, BC024928, 1097 bp	1027	862	1697
Mus musculus cDNA clone, BC031432, 1109 bp	1032	862	1697
Mus musculus cDNA clone, BC003895, 1042 bp	1035	860	1689
Homo sapiens TGFA, M31172, 867 bp	867	862	1687
Homo sapiens TGFA, BC005308, 1254 bp	889	739	1677
Pongo pygmaeus (orangutan), CR859636, 4135 bp	4019	831	1515
Synthetic construct TGFA, AY707639, 600 bp	634	492	952
Homo sapiens TGFA, AF149096, 782 bp	811	470	952

The local alignment results of two DNA sequences are derived from the real parts of the products during the cross-correlation operation. A 2-D pattern is generated by the use of Eq.(9).Figure 7(a) show the local alignment result for human and mouse TGFA genes. During the cross-correlation operation, the product of two quaternion numbers is not zero only when two sequences overlap. The non-zero products form a parallelogram in the rectangular pattern. A horizontal line appears at the index of correlation ξ =867, which corresponds to the peak correlation result shown in Fig. 7(b).In addition to the cross-correlation values, which represent the global matching information of two sequences,

the local matching information can be observed and measured in the parallelogram. To demonstrate the capability of the proposed quaternion correlation method, Figs. 7(c), 7(e), and 7(g) show the alignment results when 70 bp deletion, insertion, and substitution occur in one of the two sequences. For the cases of deletion and insertion, the horizontal line breaks into two parts, which are shifted horizontally by 70 bp. For the case of substitution, part of the line corresponding to the substituted nucleotides disappears. Figs. 7(d), 7(f), and 7(g) show that the corresponding correlation peaks appear accordingly.In addition to the peak values, the local matching positions can also be directly observed from the 2-D pattern. Compared with the 1-D correlation result, obviously, the 2-D pattern provides more information on the local matching result.

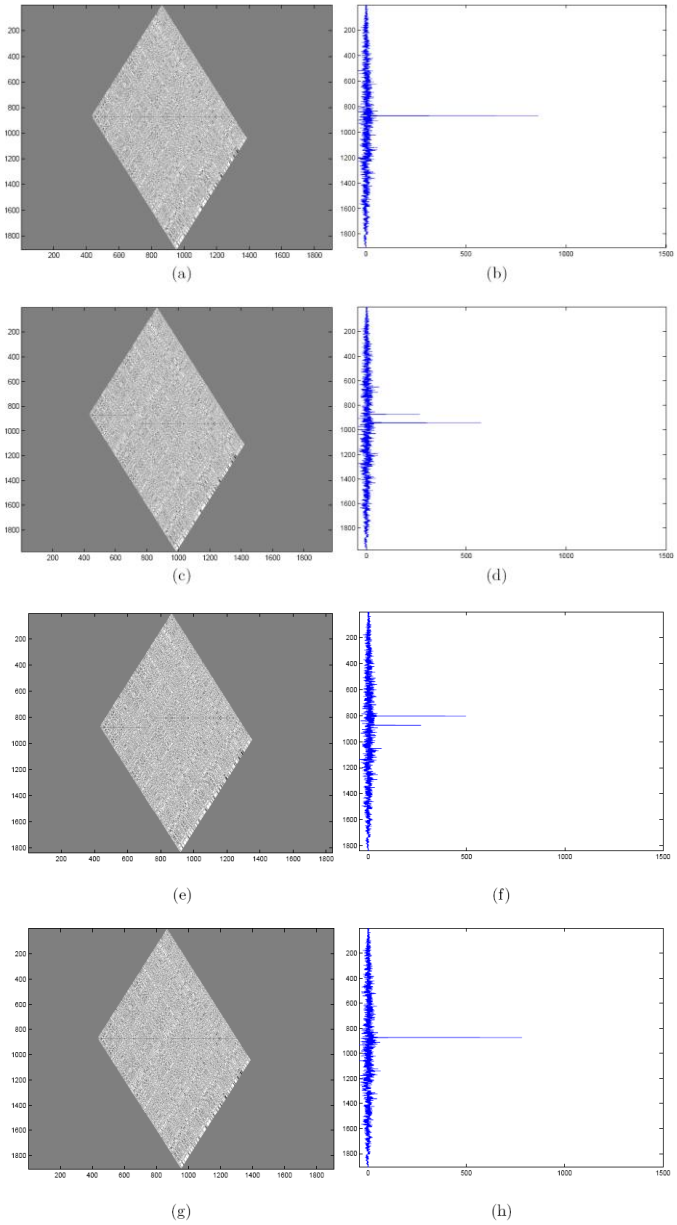

Figure 7. (a) Local matching results obtained from the cross-correlation operation of two quaternion sequences; (b) Corresponding 1-D cross-correlation result of the 2-D pattern shown in (a);(c) 70 bp deletion in one of the sequences;(d) Corresponding 1-D cross-correlation result of the 2-D pattern shown in (c).(e) 70 bp insertion in one of the sequences;(f) Corresponding 1-D cross-correlation result of the 2-D pattern shown in (e);(g) 70 bp substitution in one of the sequences;(h) Corresponding 1-D cross-correlation result of the 2-D pattern shown in (g).

Finally, Fig. 8 shows the 2-D color pattern in which the mismatching information can be directly observed. According to Fig. 8, four kinds of mismatching (AT, TA, CG, and GC) are the major parts. More information can be observed by examining the detailed parts in this pattern.

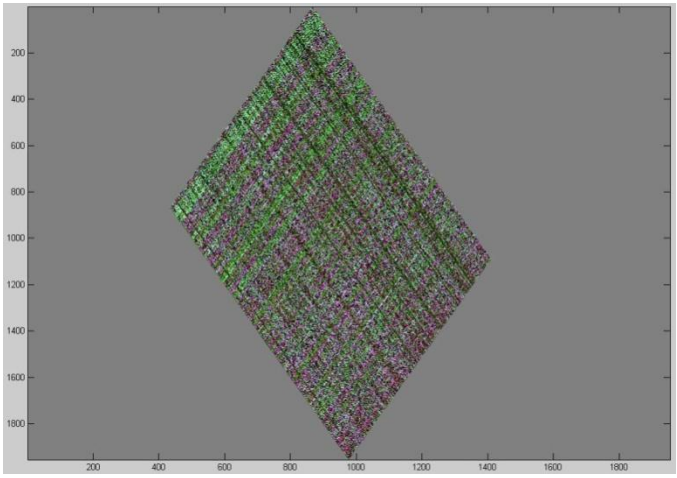

Figure 8. The 2-D pattern corresponding to the imaginary parts of quaternion correlation results of the two sequences for the Human TGFA and Mouse DHFR genes.

IV. CONCLUSION

In this study, the quaternion correlation based on quaternion number systems is proposed for DNA sequence representation and alignment. From the cross-correlation result of quaternion-number sequences, two DNA sequences can be compared in a pair-wise mode. The peak value of real part of the correlation result corresponds to the globally best-matching position of two similar sequences. On the other hand, the 2-D image obtained from the product terms in the cross-correlation operation can provide more information on local alignment of two DNA sequences. For the deletion or insertion happening in the sequences, they can be discriminated by analyzing the correlation results. Moreover, a color 2-D image can also be generated to visualize the mismatching conditions of two DNA sequences. The simulation results show that the proposed method is of promising potential in bioinformatics. Future work will focus on extracting more information and relationships between two sequences from the generated 2-D pattern.

ACKNOWLEDGMENT

This work is partly supported by National Science Council, Taiwan, under the contract number NSC 100-2628-E-224-002-MY2.

REFERENCES

[1] C. Zhang and A.K. Wong, “A genetic algorithm for multiple molecular sequence alignment,”Comput. Appl. Biosci., vol. 13, pp. 565–581, 1997.

[2] W. Choe, O.K. Ersoy, and M. Bina, “Neural network schemes for detecting rare events in human genomic DNA,” Bioinformatics, vol. 16, pp. 1062–1072, 2000.

[3] L. Hirschman, J.C. Park, J. Tsujii, L. Wong, and C.H. Wu, "Accomplishments and challenges in literature data mining for biology," Bioinformatics, vol. 18, pp. 1553–1561, 2002.

[4] P.D. Cristea, "Large scale features in DNA genomic signals," Signal Processing, vol. 83, no. 4, pp. 871–888, 2003.

[5] http://www.expasy.org/links.html

[6] http://www.ensembl.org/index.html

[7] http://www.ncbi.nlm.nih.gov/Entrez/

[8] D. Anastassiou, "Genomic signal processing,"Signal Processing Magazine, vol. 18, pp. 8–20, 2001.

[9] J. Astola, E. Dougherty, I. Shmulevich, and I. Tabus,"Genomic signal processing," Signal Processing, vol. 83, no. 4, pp. 691–694, 2003.

[10] A.V. Oppenheim, R.W. Schafer, and J.R. Buck, Discrete-Time Signal Processing, 2nd Edition,Chapter 1, Prentice Hall International, Inc., 1999.

[11] S.-Y. Hsieh, C.-W. Huang, and H.-H. Chou, "A DNA-based graph encoding scheme with its applications to graph isomorphism problems," Applied Mathematics and Computation, vol. 203, no. 2, pp. 502–512, September 2008.

[12] S.-Y. Hsieh and M.-Y. Chen, "A DNA-based solution to the graph isomorphism problem using Adleman-Lipton model with stickers," Applied Mathematics and Computation, vol. 197, no. 2, pp. 672–686, April 2008.

[13] W. Wang and D.H. Johnson, "Computing linear transforms of symbolic signals," IEEE Transactions on Signal Processing, vol. 50, pp. 628–634, 2002.

[14] S. Tiwari, S. Ramachandran, A. Bhattacharya, S. Bhattacharya,and R. Ramaswamy, "Prediction of probable genes by Fourier analysis of genomic sequences," Computer Applications in the Biosciences, vol. 13, pp 263–270, 1997.

[15] D. Anastassiou, "Frequency-domain analysis of biomolecular sequences," Bioinformatics, vol. 16, no. 12, pp. 1073–1081, 2000.

[16] Y. Magarshak , "Quaternion representation of RNA sequencesand tertiary structures," Biosystems, vol. 30, no.1–3, pp. 21–29, 1993.

[17] T. Bulow and G. Sommer, "Hypercomplex signals– A novel extension of the analytic signal to the multidimensional case," IEEE Transactions on Signal Processing, vol. 49, no. 11, pp. 2844–2852, 2001.

[18] J.-J. Shu and L.S. Ouw, "Pairwise alignment of the DNA sequence using hypercomplex number representation," Bulletin of Mathematical Biology, vol. 66, no. 5, pp. 1423--1438, 2004.

[19] Y. Li and J.-J. Shu. "Cross-correlation of DNA sequences using hypercomplex number encoding," 2006 International Conference on Biomedical and Pharmaceutical Engineering (ICBPE 2006), pp. 453–458, 11–14 Dec. 2006.

[20] J.-J. Shu and Y. Li, "Hypercomplex cross-correlation of DNA sequences," Journal of Biological Systems, vol. 18, no. 4, pp.711--725, 2010.

[21] I.L. Kantor and A.S. Solodovnikov, Hypercomplex Number: An Elementary Introduction to Algebras, New York: Springer-Verlag, 1989.

[22] S.J. Sangwine and T.A. Ell, "Color image filter based on hypercomplex convolution," IEE Proceedings of Vision, Image Signal Processing, vol. 147, no. 2, pp. 89–93, 2000.

[23] T.A. Ell and S.J. Sangwine, "Hpyercomplex Winear-Khintchine theorem with applicationto color image correlation," IEEE International Conference on Image Processing,vol. 2, pp. 792–795, 2000.

[24] K. Ueda, S.-I. Takahashi, "Digital filters with hypercomplex coefficients," 1993 IEEE International Symposium on Circuits and Systems, vol. 1, pp. 479–482, 1993.

[25] http://www.wikipedia.org/

[26] H.T. Chang, N.W. Lo, W.C. Lu, and C.J. Kuo, "Visualization of DNA sequences by use of three-dimensional trajectory," The First Asia-Pacific Bioinformatics Conference, vol. 19, pp. 81–85, Australia, Feb. 2003.

[27] H.T. Chang, "DNA sequence visualization," Chapter 4 in Advanced Data Mining Technologies in Bioinformatics, pp. 63–84, Edited by Dr. H.-H. Hsu, Idea Group, Inc., 2006.

[28] N.-W. Lo, H.T. Chang, S.W. Xiao, and C.J. Kuo,"Global visualization of DNA sequences by use of three-dimensional trajectories," Journal of Information Science and Engineering, vol. 23, no. 6, pp. 1723–1736, Nov. 2007.

[29] S.-C. Pei, J.J. Ding, and J.H. Chang, "Efficient implementation of quaternion Fourier transform, convolution, and correlation by 2-D Complex FFT," IEEE Transactions on Signal Processing, vol. 49, no. 11, pp. 2783–2797, Nov. 2001.

[30] http://www.ncbi.nlm.nih.gov/

[31] http://www.ncbi.nlm.nih.gov/BLAST/

AUTHORS PROFILE

Hsuan T. Chang received his B.S. degree in Electronic Engineering from National Taiwan University of Science and Technology, Taiwan, in 1991, and M.S. and Ph.D. degree in Electrical Engineering (EE) from National Chung Cheng University (NCCU), Taiwan, in 1993 and 1997, respectively. He was a visiting researcher at Laboratory for Excellence in Optical Data Processing, Department of Electrical and Computer Engineering, Carnegie Mellon University, from 1995 to 1996. Dr. Chang was an assistant professor in Department of Electronic Engineering in Chien-Kuo University of Technology, Changhua, Taiwan, from 1997 to 1999, an assistant professor in Department of Information Management, Chaoyang University of Technology, Wufeng, Taiwan, from 1999 to 2001, and an assistant professor and associate professor in EE Department of National Yunlin University of Science and Technology (YunTech), Douliu, Taiwan, fron 2001 to 2002 and 2003 to 2006, respectively. Her served as Chairman of Graduate Institute of Communications Engineering of Yuntech from 2008~2011. He currently is a full professor in EE Department of YunTech, Chairman of Graduate School of Engineering Technology and Science, and Deputy Dean of College of Engineering, YunTech. He was also an adjunct assistant professor in Graduate Institute of Communications Engineering of NCCU from 2000 to 2003. Dr. Chang was a visiting scholar in Institute of Information Science, Academia Sinica, Taiwan and in EE department, University of Washington, Seattle USA from 2003/7 to 2003/9 and 2007/8 to 2008/3, respectively. Dr. Chang's interests include image/video analysis, optical information processing/computing, medical image processing, and human computer interface. He has published more than 180 journal and conference papers in the above research areas. He was the recipient of the visiting research fellowship from Academia Sinica, Taiwan in 2003, and the excellent research award for young faculty in NYUST in 2005. He also received Outstanding Paper Award from 2009 MATLAB/Simulink Tech Forum Call for Papers in 2009, Taiwan. He served as the reviewer of several international journals.He served as the conference chair of 2005 Workshop on Consumer Electronics and Signal Processing held in Taiwan and was an invited speaker, session chair, and program committee in various domestic and international conferences. Dr. Chang is Senior Member of Institute of Electrical and Electronics Engineers (IEEE), Senior Member of Optical Society of America (OSA), International Society for Optical Engineering (SPIE), a member of International Who's Who (IWW), Taiwanese Association of Consumer Electronics (TACE), Asia-Pacific Signal and Information Processing Association (APSIPA), and The Chinese Image Processing and Pattern Recognition (IPPR) Society.

Chung J. Kuoreceived BS and MS degree in Power Mechanical Engineering from National TsingHua University, Taiwan, in 1982 and 1984, respectively, and PhD degree in Electrical Engineering (EE) from Michigan State University (MSU) in 1990. He joined EE Depart-ment of National Chung Cheng University (NCCU) in 1990 as an associate professor and then became a full professor in 1996. He was the chairman of Graduate Institute of Communica-tions Engineering of NCCU between 1999 and 2002. Dr. Kuo was a visiting scientist at Opto-Electronics & System Lab, Industrial Technology Research Institute in 1991 and IBM T.J. Watson Research Center from 1997 to 1998 and a consultant to several international/local companies. He was the Director of RD Center of Components Business Group (CPBG), Delta Electronics, Inc. from 2003 to 2004. In 2004, Dr. Kuo became the Senior Director of Magnetics and Microwave Business Unit of CPBG, Delta Electronics, Inc. Dr. Kuo currently is consultants of two private companies. Dr. Kuo interests in image/video signal processing, VLSI signal processing, and photonics. He is the co-director of the Signal and Media (SAM) Labs., NCCU. Dr. Kuo received the Distinguished Research Award from National Chung Cheng University in 1998, Overseas Research Fellowship from National Science Council (NSC) in 1997, Outstanding Research Award from College of Engineering, NCCU in 1997, Medal of Honor from NCCU in 1995, Research Award from NSC for consecutive 11 times, EE Fellowship from MSU in 1989, and Outstanding Academic Achievement Award from MSU in

1987. He was a guest editor for three special sections of Optical Engineering and 3D Holographic Imaging (to be published by John, Wiley and Sons) and an invited speaker and program committee chairman/member for several international/local conferences. He also serves as an Associate Editor of IEEE Signal Processing Magazine and President of SPIE Taiwan Chapter (1998-2000). Dr. Kuo is a senior member of IEEE and a member of Phi Kappa Phi, Phi Beta Delta, OSA, and SPIE and listed in Who's Who in the World.

Neng-Wen Lo received his bachelor's degree in Physics from the Tunghai University in 1986 and Ph.D. degree in Biochemistry from the State University of New York at Buffalo, USA, in 1997. He was a research fellow at the Oncology Center of Johns Hopkins School of Medicine in Maryland, USA, from 1997 to 1999. In 1999, he returned to Taiwan and joined the Department of Animal Science and Biotechnology at Tunghai University. He is currently an Associate Professor and Head of the Agriculture Extension Center, College of Agriculture. Dr. Lo's research interest is in Computational Biology and in Reproduction Biology. He has published more than 40 journal and conference papers. He was the chief editor of Tunghai Journal in 2003 and ever served as referees for several journals. Dr. Lo is a founding member of the Bioinformatics Society in Taiwan. He is also a member of the Society for the Study of Reproduction and a permanent member of Chinese Society of Animal Science.

Wei-Z.Lv Biography is not available.