Journal of Biological Systems, Vol. 18, No. 4 (2010) 711–725

DOI: 10.1142/S0218339010003470

World Scientific
www.worldscientific.com

HYPERCOMPLEX CROSS-CORRELATION OF DNA SEQUENCES

JIAN-JUN SHU* and YAJING LI

School of Mechanical and Aerospace Engineering
Nanyang Technological University
50 Nanyang Avenue, Singapore 639798
**mjjshu@ntu.edu.sg*

Received 16 October 2009
Accepted 21 December 2009

A hypercomplex representation of DNA is proposed to facilitate comparing DNA sequences with fuzzy composition. With the hypercomplex number representation, the conventional sequence analysis method, such as, dot matrix analysis, dynamic programming, and cross-correlation method have been extended and improved to align DNA sequences with fuzzy composition. The hypercomplex dot matrix analysis can provide more control over the degree of alignment desired. A new scoring system has been proposed to accommodate the hypercomplex number representation of DNA and integrated with dynamic programming alignment method. By using hypercomplex cross-correlation, the match and mismatch alignment information between two aligned DNA sequences are separately stored in the resultant real part and imaginary parts respectively. The mismatch alignment information is very useful to refine consensus sequence based motif scanning.

Keywords: Hypercomplex Representation; DNA Alignment.

1. Introduction

DNA molecules are composed of four linearly linked nucleotides, adenine (A), thymine (T), guanine (G) and cytosine (C). A DNA sequence can be represented as a permutation of four characters A, T, G, and C at different lengths. The standard symbolic representation of DNA sequence has definite advantages in what concerns storage, search and retrieval of genomic information, but has some limitation in handling and processing genomic information for pattern matching. Converting DNA sequence into numerical sequence opens the possibility of applying signal processing methods to the analysis of genomic data. There have been several attempts to attach numerical values to the symbolic DNA sequences for different purposes.[1–7]

Cross-correlation is an engineering technique for comparing two numerical signals. When comparing two DNA sequences, two kinds of numerical representation methods were used in the literature: One is 4-vector encoding by which a DNA

*Corresponding author.

sequence is decomposed into 4 independent binary indicator sequences[1,8]; Another is complex number encoding by which a DNA sequence is converted into one complex number sequence (A, T, G and C are mapped to 1, -1, i and $-i$ respectively).[2] With the 4-vector encoding, four cross-correlations are needed to get the total match count between a pair of DNA sequences at each alignment. The disadvantage of this method is that all the mismatch alignment information between the two DNA sequences is missed. With the complex number encoding, only one cross-correlation is needed to get the match and mismatch alignment information. However, the real part of the cross-correlation equals the number of matches at each alignment of the two DNA sequences minus the number of complementary mismatch (alignment between A and T as well as alignment between G and C). So the real part of cross-correlation is actually an approximate measure of the similarity between the pair of DNA sequences. The peak in the real part, which indicates the similarity between the pair of DNA sequences, may disappear due to the offset of the complementary mismatch to the match. On the other hand, complex number representation cannot encode the DNA base codes with fuzzy composition, such as IUPAC codes[9] for representation of degenerate consensus sequences. This disadvantage limits the use of cross-correlation to deal with the problem of DNA consensus sequence matching.

The hypercomplex numbers form a non-commutative four-dimensional number system that extends the commutative one-dimensional real and two-dimensional complex numbers. Since there are four types of nucleotide, a four-dimensional space is essential to represent the DNA codes fully. The hypercomplex number representation of DNA base code proposed in[10] can take the occurrence frequency of each nucleotide in a DNA nucleotide base code into full account. It is very suitable for numerical encoding DNA base code with fuzzy composition. A hypercomplex cross-correlation method in the DNA consensus sequence matching analysis is proposed in this paper by using the hypercomplex number encoding of DNA sequences.

2. Hypercomplex Representation of DNA Motif

DNA motifs are short conserved domains that are presumed to have biological functions. Often they indicate sequence-specific binding sites for proteins such as nucleases and transcription factors. In DNA binding sites, only some positions are conserved among all the sites, and the other positions have a range of variability. So a DNA binding site is usually represented by a consensus sequence or a position-weight matrix (PWM). As an example, TATA-box consensus sequence, base frequency and PWM for eukaryotic RNA polymerase II promoters are listed in Table 1.

The consensus sequence 'TATAWAWR' can be represented by the following three different hypercomplex vectors:

(1) Consensus sequence based hypercomplex representation of TATA box:

$$\mathbf{V}_1 = (\mathbf{i}, 1, \mathbf{i}, 1, 0.5 + 0.5\mathbf{i}, 1, 0.5 + 0.5\mathbf{i}, 0.5 + 0.5\mathbf{j}),$$

Table 1. TATA-box consensus sequence, base frequency and PWM for eukaryotic RNA polymerase II promoters,[11] and their hypercomplex representation.

Consensus	T	A	T	A	W	A	W	R
$P_{\mathrm{A}}(\%)$	4.1	90.5	0.8	91.0	68.9	92.5	57.1	39.8
$P_{\mathrm{T}}(\%)$	79.5	9.0	96.1	7.7	31.1	1.6	31.1	8.5
$P_{\mathrm{G}}(\%)$	4.6	0.5	0.5	1.3	0.0	5.1	11.3	40.4
$P_{\mathrm{C}}(\%)$	11.8	0.0	2.6	0.0	0.0	0.8	0.5	11.3
Position	1	2	3	4	5	6	7	8
W_{A}	−3.05	0	−4.61	0	0	0	0	−0.01
W_{T}	0	−2.28	0	−2.34	−0.52	−3.65	−0.37	−1.4
W_{G}	−2.74	−4.28	−4.61	−3.77	−4.73	−2.65	−1.5	0
W_{C}	−2.06	−5.22	−3.49	−5.17	−4.63	−4.12	−3.74	−1.13

based on the hypercomplex representation of DNA base code[10];

(2) Frequency based hypercomplex representation of TATA box:

$$\mathbf{V}_2 = \begin{pmatrix} 0.041 + 0.795\mathbf{i} + 0.046\mathbf{j} + 0.118\mathbf{k}, & 0.905 + 0.09\mathbf{i} + 0.005\mathbf{j}, \\ 0.008 + 0.961\mathbf{i} + 0.005\mathbf{j} + 0.026\mathbf{k}, & 0.91 + 0.077\mathbf{i} + 0.013\mathbf{j}, \\ 0.689 + 0.311\mathbf{i}, & 0.925 + 0.016\mathbf{i} + 0.051\mathbf{j} + 0.008\mathbf{k}, \\ 0.571 + 0.311\mathbf{i} + 0.113\mathbf{j} + 0.005\mathbf{k}, & 0.398 + 0.085\mathbf{i} + 0.404\mathbf{j} + 0.113\mathbf{k} \end{pmatrix},$$

based on every column of the base frequency, $P_A + P_T\mathbf{i} + P_G\mathbf{j} + P_C\mathbf{k}$, in Table 1;

(3) PWM based hypercomplex representation of TATA box:

$$\mathbf{V}_3 = \begin{pmatrix} -3.05 - 2.74\mathbf{j} - 2.06\mathbf{k}, & -2.28\mathbf{i} - 4.28\mathbf{j} - 5.22\mathbf{k}, \\ -4.61 - 4.61\mathbf{j} - 3.49\mathbf{k}, & -2.34\mathbf{i} - 3.77\mathbf{j} - 5.17\mathbf{k}, \\ -0.52\mathbf{i} - 4.73\mathbf{j} - 4.63\mathbf{k}, & -3.65\mathbf{i} - 2.65\mathbf{j} - 4.12\mathbf{k}, \\ -0.37\mathbf{i} - 1.5\mathbf{j} - 3.74\mathbf{k}, & -0.01 - 1.4\mathbf{i} - 1.13\mathbf{k} \end{pmatrix}$$

based on every column of the hypercomplex position-weight vector (HPWV), $W_A + W_T\mathbf{i} + W_G\mathbf{j} + W_C\mathbf{k}$, in Table 1.

3. Hypercomplex Cross-Correlation

The cross-correlation is a measure of similarity between two waveforms as a function of a time-lag applied to one of them. It is commonly used to search a long duration signal for a shorter. The cross-correlation is similar in nature to the convolution of two functions. The mathematical form of cross-correlation function can be generalized by using a hypercomplex presentation. Here consider $\mathbf{u} = (u_0, u_1, \ldots, u_{N-1})$ and $\mathbf{v} = (v_0, v_1, \ldots, v_{N-1})$, two hypercomplex vectors of length N. The hypercomplex cyclic cross-correlation operation of $\mathbf{u}$ and $\mathbf{v}$ is defined as the hypercomplex vector $\mathbf{w} = (w_0, w_1, \ldots, w_{N-1})$:

$$\mathbf{W}_l = (\mathbf{u} * \bar{\mathbf{v}})_l = \sum_{n=0}^{N-1} u_{(n+l) \mod N}\, \bar{v}_n \quad l = 0,\ 1, \ldots, N-1 \tag{1}$$

where $\bar{v}_n$ is the conjugate of v_n. The shift operation on u_n is carried out cyclically using modulo arithmetic for the subtraction.

3.1. *Hypercomplex pairwise DNA sequence alignment*

Hypercomplex cross-correlation is an extension of cross-correlation concept to hypercomplex vector. So the process of using hypercomplex cross-correlation to compare DNA sequences is similar to the one used in Ref. 2. The difference is that the hypercomplex number is used to encode DNA base code: 1 for A, **i** for T, **j** for G and **k** for C.

The algorithm for pairwise DNA sequence alignment with hypercomplex cross-correlation is as follows, and the schematic graph is shown in Fig. 1.

Algorithm 1: Given two DNA sequences X and Y of length N and M respectively, calculate pairwise alignment between X and Y with hypercomplex cross-correlation:

Step 1: Transform the two DNA sequences X and Y into hypercomplex vectors $(x_0, x_1, \ldots, x_{N-1})$ and $(y_0, y_1, \ldots, y_{M-1})$.

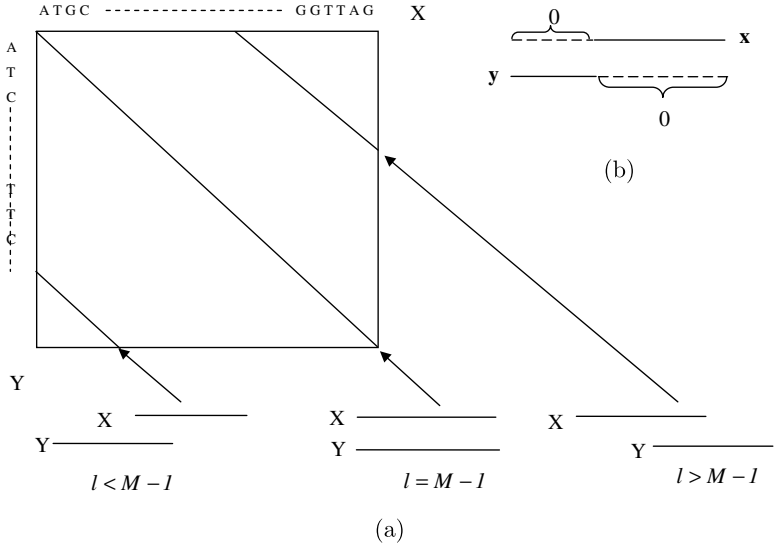


Fig. 1. Sketch for computing sequence similarity by hypercomplex cross-correlation: (a) X and Y are two DNA sequences; $l - M + 1$ is the relative position between X and Y, the position is set when x_0 are aligned with y_0 as $M - 1$ for l; the real part of hypercomplex cross-correlation is actually equal to the match count of each diagonal. (b) After transforming to hypercomplex vector, pad $\mathbf{x}$ on the left with $M - 1$ 0's and $\mathbf{y}$ on the right with $N - 1$ 0's to form two vectors

$$\mathbf{x} = \left(\overbrace{0, \ldots, 0}^{M-1}, x_0, x_1, \ldots, x_{N-1} \right) \text{ and } \mathbf{y} = \left(y_0, y_1, \ldots, y_{M-1}, \underbrace{0, \ldots, 0}_{N-1} \right)$$

of length $N + M - 1$.

Step 2: Pad $(x_0, x_1, \ldots, x_{N-1})$ on the left with $M-1$ 0's and $(y_0, y_1, \ldots, y_{M-1})$ on the right with $N-1$ 0's to form two vectors

$$\mathbf{x} = \left(\overbrace{0, \ldots, 0}^{M-1}, x_0, x_1, \ldots, x_{N-1} \right) \text{ and } \mathbf{y} = \left(y_0, y_1, \ldots, y_{M-1}, \underbrace{0, \ldots, 0}_{N-1} \right)$$

of length N + M − 1.

Step 3: Calculate the hypercomplex cyclic cross-correlation $\mathbf{z} = \mathbf{x} * \bar{\mathbf{y}}$ by Eq. (1).

Step 4: Plot the real part $\mathbf{r} = (r_0, r_1, \ldots, r_{N+M-1})$ of $\mathbf{z}$ against the shift l. A peak in the real part of hypercomplex cross-correlation indicates similarity between two DNA sequences. The relative position $l-M+1$ corresponding to the maximum of $\mathbf{r}$ is the position that corresponds to the maximal similarity.

When cross-correlation is performed between two hypercomplex represented DNA sequences, according to the hypercomplex algebra, the hypercomplex product in Eq. (1) results in 1 if the two aligned bases are matched, and $\pm\mathbf{i}$ or $\pm\mathbf{j}$ or $\pm\mathbf{k}$ if they are mismatched (for the detailed results, see Table 2), here $\pm\mathbf{i}$ corresponding to the complementary mismatch (A *vs.* T or G *vs.* C); $\pm\mathbf{j}$ to the transitional mismatch (A *vs.* G or T *vs.* C); and $\pm\mathbf{k}$ to the transversal mismatch (A *vs.* C or G *vs.* T).

In summary, we have:

$$\begin{cases} d = n_{\mathrm{AA}} + n_{\mathrm{TT}} + n_{\mathrm{GG}} + n_{\mathrm{CC}} \\ a = n_{\mathrm{TA}} + n_{\mathrm{CG}} - n_{\mathrm{AT}} - n_{\mathrm{GC}} \\ b = n_{\mathrm{GA}} + n_{\mathrm{TC}} - n_{\mathrm{AG}} - n_{\mathrm{CT}} \\ c = n_{\mathrm{CA}} + n_{\mathrm{GT}} - n_{\mathrm{AC}} - n_{\mathrm{TG}} \end{cases} \tag{2}$$

where d, a, b and c denote the real and three imaginary parts of hypercomplex cross-correlation $\mathbf{z}$, n_{XY} denotes the number of aligned base pair XY(X, Y ∈ A, T, G, C). Thus, the real part of the hypercomplex cross-correlation gives the number of matches at each alignment of the two DNA sequences, the imaginary part reflects the mismatch alignment information: Every complementary mismatch TA or CG increases **i** imaginary part a by one, due to the non-commutation of hypercomplex product, every mismatch AT or GC decreases a by one; similarly every transitional mismatch GA or TC increase **j** imaginary part b by one, every mismatch AG or CT decreases b by one; every transversal mismatch CA or GT increases **k** imaginary part c by one, every mismatch AC or TG decreases c by one. Thus, a peak in the real part of hypercomplex cross-correlation indicates similarity between two DNA

Table 2. Hypercomplex product results between hypercomplex represented DNA base codes.

	A	T	G	C
A	1	$-\mathbf{i}$	$-\mathbf{j}$	$-\mathbf{k}$
T	**i**	1	$-\mathbf{k}$	**j**
G	**j**	**k**	1	$-\mathbf{i}$
C	**k**	$-\mathbf{j}$	**i**	1

sequences. Furthermore, because the real part of hypercomplex cross-correlation only reflects the match alignment information, the peak value can be used as a measure of similarity between two DNA sequences: The larger the peak, the more similar the two DNA sequences.

3.2. *Hypercomplex consensus sequence based motif scanning*

When the vector $\mathbf{y}$ is the hypercomplex representation of consensus sequence, the hypercomplex cross-correlation can be used to do consensus sequence based motif scanning. If the two aligned DNA bases have hypercomplex representations $\mathbf{x} = P_{\mathrm{A}}^{\mathrm{x}} + P_{\mathrm{T}}^{\mathrm{x}}\mathbf{i} + P_{\mathrm{G}}^{\mathrm{x}}\mathbf{j} + P_{\mathrm{C}}^{\mathrm{x}}\mathbf{k}$ and $\mathbf{y} = P_{\mathrm{A}}^{\mathrm{y}} + P_{\mathrm{T}}^{\mathrm{y}}\mathbf{i} + P_{\mathrm{G}}^{\mathrm{y}}\mathbf{j} + P_{\mathrm{C}}^{\mathrm{y}}\mathbf{k}$, where P is the base frequency, the multiplication in Eq. (1) is denoted as R_{xy} and calculated by:

$$\begin{aligned} R_{\mathrm{xy}} = \mathbf{x}\bar{\mathbf{y}} &= P_{\mathrm{A}}^{\mathrm{x}}P_{\mathrm{A}}^{\mathrm{y}} + P_{\mathrm{T}}^{\mathrm{x}}P_{\mathrm{T}}^{\mathrm{y}} + P_{\mathrm{G}}^{\mathrm{x}}P_{\mathrm{G}}^{\mathrm{y}} + P_{\mathrm{C}}^{\mathrm{x}}P_{\mathrm{C}}^{\mathrm{y}} \\ &\quad + (P_{\mathrm{T}}^{\mathrm{x}}P_{\mathrm{A}}^{\mathrm{y}} + P_{\mathrm{C}}^{\mathrm{x}}P_{\mathrm{G}}^{\mathrm{y}} - P_{\mathrm{A}}^{\mathrm{x}}P_{\mathrm{T}}^{\mathrm{y}} - P_{\mathrm{G}}^{\mathrm{x}}P_{\mathrm{C}}^{\mathrm{y}})\mathbf{i} \\ &\quad + (P_{\mathrm{G}}^{\mathrm{x}}P_{\mathrm{A}}^{\mathrm{y}} + P_{\mathrm{T}}^{\mathrm{x}}P_{\mathrm{C}}^{\mathrm{y}} - P_{\mathrm{A}}^{\mathrm{x}}P_{\mathrm{G}}^{\mathrm{y}} - P_{\mathrm{C}}^{\mathrm{x}}P_{\mathrm{T}}^{\mathrm{y}})\mathbf{j} \\ &\quad + (P_{\mathrm{C}}^{\mathrm{x}}P_{\mathrm{A}}^{\mathrm{y}} + P_{\mathrm{G}}^{\mathrm{x}}P_{\mathrm{T}}^{\mathrm{y}} - P_{\mathrm{A}}^{\mathrm{x}}P_{\mathrm{C}}^{\mathrm{y}} - P_{\mathrm{T}}^{\mathrm{x}}P_{\mathrm{G}}^{\mathrm{y}})\mathbf{k} \end{aligned} \tag{3}$$

The real part of R_{xy} is the probability of finding a match in the alignment since the hypercomplex number representation assigned to each of the DNA nucleotide base code is based on the base frequency P that each base appears in the base code. The imaginary parts of R_{xy} reflect the probability of finding a mismatch in the alignment: the $\mathbf{i}$ imaginary part is the probability of finding TA or CG pair minus the probability of finding AT or GC pair; the $\mathbf{j}$ imaginary part is the probability of finding GA or TC pair minus the probability of finding AG or CT pair; the $\mathbf{k}$ imaginary part is the probability of finding CA or GT pair minus the probability of finding AC or TG pair. Thus, the real part of the hypercomplex cross-correlation $\mathbf{z}$ gives a score that can be used to measure the similarity between DNA substring and the consensus sequence.

The algorithm for DNA motif scanning with hypercomplex cross-correlation is as follows:

Algorithm 2: Given a consensus sequence Y of length M and a DNA sequence X of length N, scan DNA sequence X to identify all the similar segment of X to the consensus sequence Y.

Step 1: Transform the two DNA sequences X and Y into hypercomplex vectors $(x_0, x_1, \ldots, x_{N-1})$ and $(y_0, y_1, \ldots, y_{M-1})$.

Step 2: Pad $(x_0, x_1, \ldots, x_{N-1})$ on the left with $M-1$ 0's and $(y_0, y_1, \ldots, y_{M-1})$ on the right with $N-1$ 0's to form two vectors

$$\mathbf{x} = \left(\overbrace{0, \ldots, 0}^{M-1}, x_0, x_1, \ldots, x_{N-1}\right) \quad \text{and} \quad \mathbf{y} = \left(y_0, y_1, \ldots, y_{M-1}, \underbrace{0, \ldots, 0}_{N-1}\right)$$

of length $N + M - 1$.

Step 3: Calculate the hypercomplex cyclic cross-correlation $\mathbf{z} = \mathbf{x} * \bar{\mathbf{y}}$ by Eq. (1).
Step 4: Plot the real part $\mathbf{r} = (r_M, r_{M+1}, \ldots, r_N)$ of $\mathbf{z}$ against l. The position $l - M + 1$ corresponding to the maximum of $\mathbf{r}$ is the start position of the match between DNA substring and the consensus sequence (the potential binding site).

The hypercomplex cross-correlation can be implemented by direct evaluation of the summation formula in Eq. (1), which requires running time $O(N^2)$. The cost of performing direct evaluation of cross-correlation can be reduced to $O(N \ln N)$ by the hypercomplex Fourier transform.

3.3. *Hypercomplex PWM based motif scanning*

If the vector $\mathbf{y}$ is changed to the hypercomplex representation of PWM, the hypercomplex cross-correlation can be used to do profile matching.[12,13] The real part of multiplication in Eq. (1) is $W_{\mathrm{A}}^{\mathrm{x}} W_{\mathrm{A}}^{\mathrm{y}} + W_{\mathrm{T}}^{\mathrm{x}} W_{\mathrm{T}}^{\mathrm{y}} + W_{\mathrm{G}}^{\mathrm{x}} W_{\mathrm{G}}^{\mathrm{y}} + W_{\mathrm{C}}^{\mathrm{x}} W_{\mathrm{C}}^{\mathrm{y}}$. Thus, the real part of the hypercomplex cross-correlation $\mathbf{z}$ equals to the profile match score.

3.4. *Hypercomplex Fourier transform (HFT)*

A hypercomplex number can be represented in the polar form by generalizing the Euler's formula for complex numbers. That is:

$$\mathbf{z} = d + a\mathbf{i} + b\mathbf{j} + c\mathbf{k} = |\mathbf{z}| e^{\boldsymbol{\mu}\theta} = |\mathbf{z}|(\cos\theta + \boldsymbol{\mu}\sin\theta) \tag{4}$$

where

$$\cos\theta = \frac{d}{|\mathbf{z}|}, \quad \sin\theta = \frac{\sqrt{a^2 + b^2 + c^2}}{|\mathbf{z}|}, \quad |\mathbf{z}| = \sqrt{d^2 + a^2 + b^2 + c^2} \tag{5}$$

then $\boldsymbol{\mu}$ can be expressed as

$$\boldsymbol{\mu} = \frac{a}{\sqrt{a^2 + b^2 + c^2}}\mathbf{i} + \frac{b}{\sqrt{a^2 + b^2 + c^2}}\mathbf{j} + \frac{c}{\sqrt{a^2 + b^2 + c^2}}\mathbf{k} \tag{6}$$

$$\boldsymbol{\mu}^2 = -1$$

$\boldsymbol{\mu}$ is a unit pure hypercomplex number, and it is referred to the eigen-axis. Note that $\boldsymbol{\mu}$ represents the direction in the 3-D space of imaginary part of hypercomplex number and θ is referred to the eigen-angle.

The discrete HFT is generalized from discrete complex Fourier transform (FT). Due to the non-commutativity of hypercomplex number multiplication, HFT is defined as:

$$Z(\kappa) = \mathrm{HFT}\{z_n\} = \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} z_n \mathrm{e}^{-\boldsymbol{\mu}_0 \frac{2\pi}{N} \kappa n} \tag{7}$$

Its inverse discrete hypercomplex Fourier transform (IHFT) is given by

$$z_n = \mathrm{IHFT}\{Z(\kappa)\} = \frac{1}{\sqrt{N}} \sum_{\kappa=0}^{N-1} Z(\kappa) \mathrm{e}^{\boldsymbol{\mu}_0 \frac{2\pi}{N} \kappa n} \tag{8}$$

where the vector $\boldsymbol{\mu}_0$ is called the axis of the transform, and it is an arbitrary unit pure hypercomplex number. For computing convenience, the vector $\mathbf{i}$ is chosen as the transform axis here after.

3.5. *Implementation of hypercomplex Fourier transform (HFT) by Fourier transform (FT) and inverse Fourier transform (IFT)*

Assume the input hypercomplex series $z_n = d_n + a_n\mathbf{i} + b_n\mathbf{j} + c_n\mathbf{k}$ and a given transform axis is $\mathbf{i}$, HFT $Z(\kappa)$ is calculated as:

$$Z(\kappa) = \mathrm{HFT}\{z_n\} = \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} z_n \, \mathrm{e}^{-i \frac{2\pi}{N} \kappa n} = \mathrm{FT}\{z_n^{(1)}\} + \mathrm{IFT}\{z_n^{(2)}\}\,\mathbf{j} \tag{9}$$

where $z_n = z_n^{(1)} + z_n^{(2)}\,\mathbf{j}$, $z_n^{(1)} = d_n + a_n\mathbf{i}$ and $z_n^{(2)} = b_n + c_n\mathbf{i}$.

Assume two hypercomplex vectors $\mathbf{x}$ and $\mathbf{y}$, hypercomplex cross-correlation can be calculated by the following equation[14]:

$$\begin{aligned} z_n &= \mathrm{IHFT}\{Z(\kappa)\} = \frac{1}{\sqrt{N}} \sum_{\kappa=0}^{N-1} Z(\kappa)\, \mathrm{e}^{\mathbf{i} \frac{2\pi}{N} \kappa\, n} \\ &= \mathrm{IHFT}\left\{\mathrm{HFT}\{x_n\}\,\overline{\mathrm{HFT}\{y_n^{(1)}\}} - \mathrm{IHFT}\{x_n\}\,\mathrm{IHFT}\{y_n^{(2)}\}\,\mathbf{j}\right\}. \end{aligned} \tag{10}$$

By using HFT, the computational complexity of the Algorithm 2 is reduced to $O(N \ln N)$.[14,15]

4. Results and Discussion

4.1. *Hypercomplex pairwise alignment of DNA*

Three human immunodeficiency virus (HIV) sequences are used to show application of hypercomplex cross-correlation for DNA pairwise alignment. There are two human type 1 isolates HIVMN (HIV type 1, isolate MN, GenBank accession number M17449) and HIVRF (HIV type 1, isolate RF, GenBank accession number M17451), and one simian virus, SIVMM142 (Simian immunodeficiency virus isolate MM142 m-83, GenBank accession number Y00277). Each sequence is between 9000 and 10,000 bases in length and available in GenBank.[16]

Figure 2(a) shows the real part of the hypercomplex cross-correlation of HIVMN with HIVRF. The horizontal axis is the relative shift of HIVRF with respect to HIVMN. The vertical axis is the number of matches. Figure 2(b) is similar to

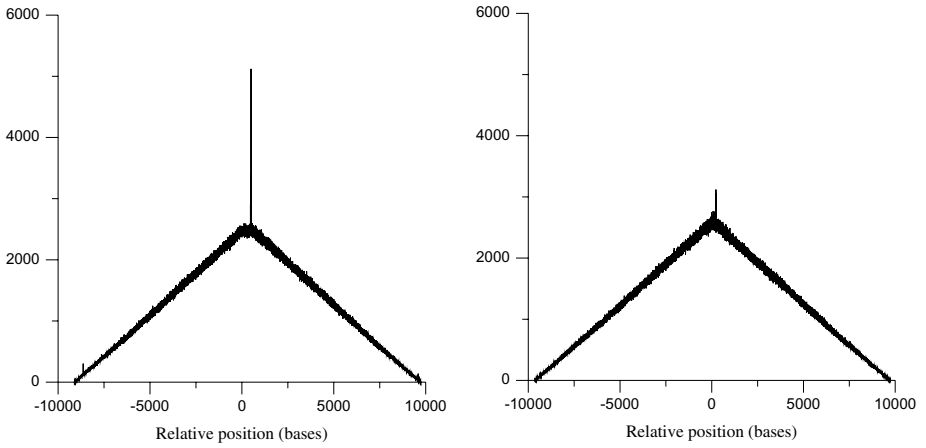


Fig. 2. Real part of hypercomplex correlation of (a) HIVMN vs. HIVRF and (b) HIVMN vs. SIVMM142.

Fig. 2(a) but shows the real part of the hypercomplex cross-correlation of HIVMN with SIVMM142.

A peak in the real part of the hypercomplex cross-correlation indicates similarity between two compared sequences. The peak value can be used to measure the similarity, for example, as expected, the peak around the 0 shift, is much larger for the comparison of two different HIV isolates than for comparison of the SIV and HIV isolates.

4.2. *Hypercomplex consensus sequence based motif scanning*

Another application of hypercomplex cross-correlation is to scan a query DNA sequence for the potential transcription binding sites. Here an example of using hypercomplex correlation is given to locate the potential TATA box in query sequence of H. sapiens H4/g gene (gi number: 32003). TATA-box consensus sequence, base frequency and PWM for eukaryotic RNA polymerase II promoters and their corresponding hypercomplex representation are listed in Table 1. From experiment, the TATA signal locates at position 185 as substring 'TATTTAAG'. When the conventional string match method[17] is used to scan H. sapiens H4/g gene for potential TATA box, there is not an exact match to the consensus sequence 'TATAWAWR'. However, when using cross-correlation with hypercomplex encoding, the potential TATA box can be found by setting reasonable score threshold.

The real part of hypercomplex cross-correlation of H. sapiens H4/g gene against the consensus sequence 'TATAWAWR' is shown in Fig. 3. If the score threshold is set as 5.5, two potential TATA boxes are identified to be located at position 185 and 381 of the query DNA sequence respectively. Then the best candidate should

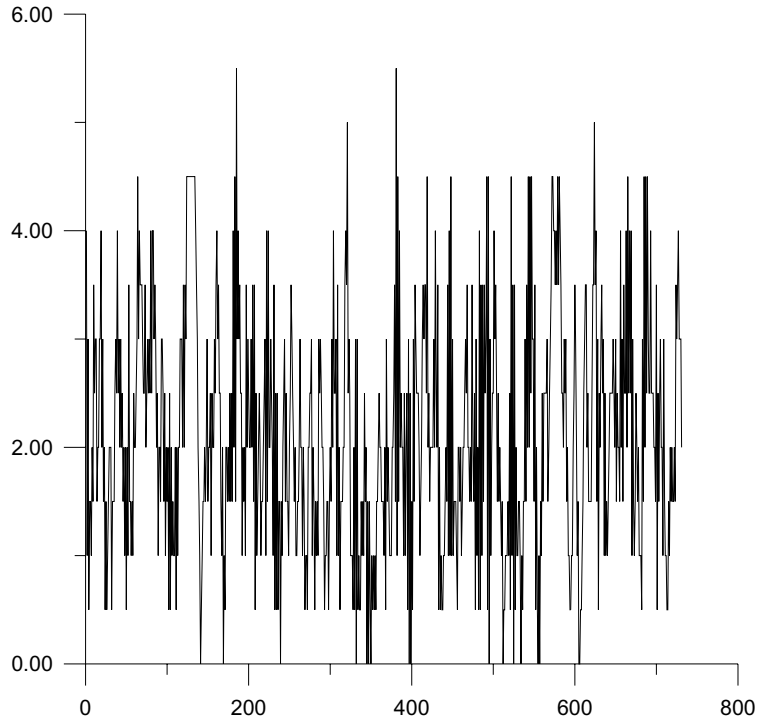


Fig. 3. Real part of hypercomplex cross-correlation of H4/g gene against consensus sequence 'TATAWAWR'. Hypercomplex representation of consensus sequence is the vector $\mathbf{V}_1$ in Table 1.

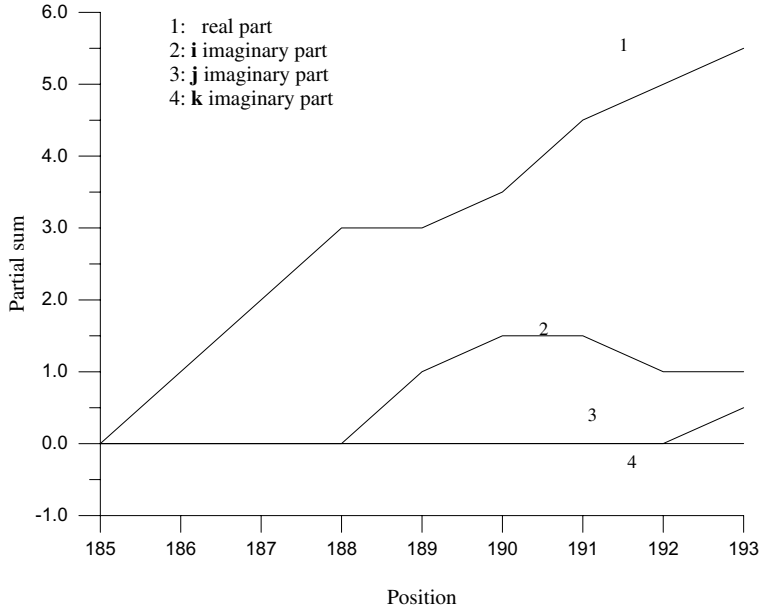
be identified to show most similar to consensus sequence. One simple way is to line the candidate substring against the consensus sequence and then perform the base-by-base comparison. This operation can be expressed in an alignment graph. For the candidate at position 185 (the candidate 1: 'TATTTAAG'), the alignment graph reads as follows:

```
T A T T  T  A A G
| | |    |  | | |
T A T A W A W R
```

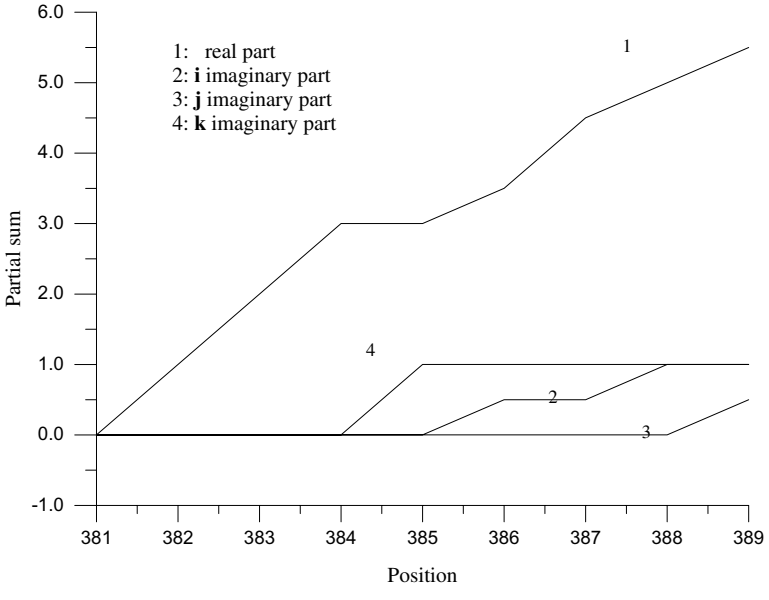
For the candidate at position 381 (the candidate 2: 'TATCTATG'), the alignment graph is:

```
T A T C  T  A T G
| | |    |  | | |
T A T A W A W R
```

From the alignment graphs, the two candidates match the consensus sequence at all the positions except for the position 4. For the candidate 1, the mismatch



(a)



(b)

Fig. 4. Partial sum plots for H. sapiens H4/g gene compared with consensus sequence 'TATAWAWR', showing the match and mismatch alignment information from positions (a) 185 and (b) 381.

is a complementary mutation of 'A to T'; for the candidate 2, the mismatch is a transversal mutation 'A to C'. Because complementary mismatches occur more probably than the transversal mismatch in the binding sites, the candidate 1 is more similar to the consensus sequence. Thus, the substring at position 185 is the best candidate.

The second approach to identify the best candidate is to use the mismatch alignment information from the hypercomplex cross-correlation. For each peak in the Fig. 2, the correlation shift value l is determined, hold l fixed and plot in Fig. 4 the partial sum function $S_l(m)$:

$$S_l(m) = \sum_{n=0}^{m-1} x_{(n+l) \mod N} \overline{y_n} \quad m = 1, 2, \ldots, N \tag{11}$$

The real part of the partial sum reflects the match alignment information and the imaginary parts reflect the mismatch alignment information. For the real part plot (line 1), a region with slope of +1 indicates a perfect match, a horizontal region corresponds to a mismatch, a region with slope less than 1 corresponds to a fuzzy match (for example the alignment between T and W). For the **i** imaginary part plot

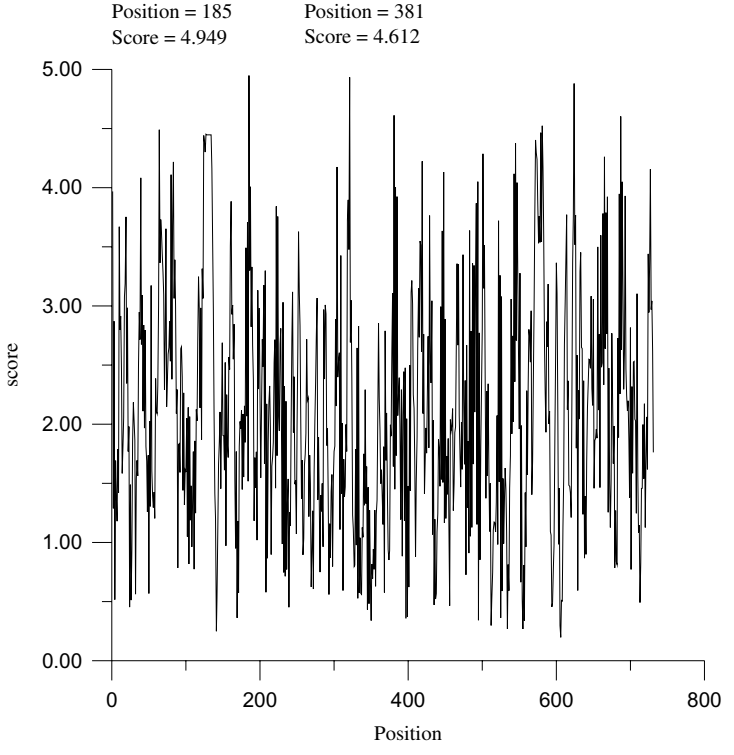


Fig. 5. Real part of hypercomplex cross-correlation of H. sapiens H4/g gene against frequency based consensus sequence. Hypercomplex representation is the vector $\mathbf{V}_2$ in Table 1.

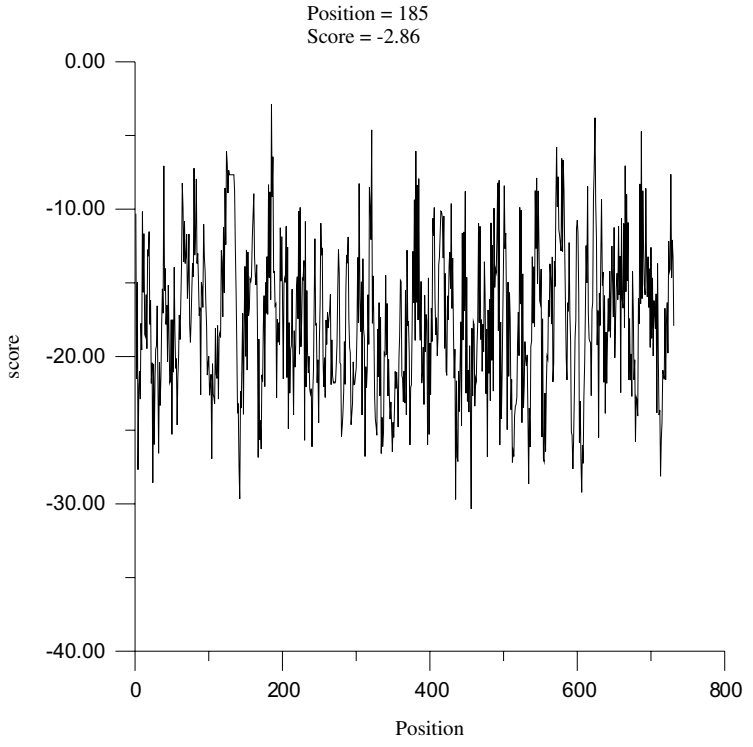


Fig. 6. Real part of hypercomplex cross-correlation of H. sapiens H4/g gene against the hypercomplex position weight vector $\mathbf{V}_3$ shown in Table 1.

(line 2), a region with slope of +1 or −1 indicates a perfect complementary mismatch, a horizontal region indicates no complementary mismatch at that region, a region with slope less than 1 corresponds to a fuzzy mismatch. The similar meaning is for **j** and **k** imaginary part plots (lines 3 and 4).

Figure 4 shows that the mismatch is located at the position 4 of each candidate substring: for the candidate 1, the mismatch is a complementary mutation; for the candidate 2, the mismatch is a transversal mutation. Thus, the candidate at position 185 is the best candidate.

Another approach to identify the best candidate is to use the hypercomplex encoding of TATA box based on the base frequency at each position (the vector $\mathbf{V}_2$ shown in Table 1). The real part of hypercomplex cross-correlation is shown in Fig. 5. The candidate at the position 185 can easily be identified as the best candidate because it corresponds to maximal score of the real part.

4.3. *Hypercomplex PWM based motif scanning*

Finally, the real part of hypercomplex correlation result of H. sapiens H4/g gene against the HPWV (the vector $\mathbf{V}_3$ shown in Table 1) is shown in Fig. 6. The real

part of the hypercomplex correlation result equals to the profile match score. The substring can easily be identified at position 185 as the most probably TATA box.

5. Conclusion

The conventional way of handling motif scanning is to search good match substring alignments, such as, the real-number-based fast Fourier transform approach[8] and the hypercomplex-number-based dynamic programming matrix approach.[10] However, these approaches, solely based on match alignment information, are susceptible to a high number of ambiguous or falsely predicted sites, especially if the motif sequences are not well conserved. In order to reduce the number of false positives, the mismatch component of the alignment is considered. With hypercomplex representation of DNA, match and mismatch alignment information between two compared DNA sequences are separated and stored in the real part and imaginary part of the cross-correlation result respectively. So the real part of cross-correlation is an exact measure of the similarity between the pair of DNA sequences at each alignment. Moreover, with the emphasis on probabilistic occurrence of each DNA base on specific position by hypercomplex representation of DNA, the cross-correlation method can be used to scan a query DNA sequence for potential transcription binding site, the real part of the hypercomplex cross-correlation gives a score that indicates the similarity degree of DNA substring and the consensus sequence. With reasonable score threshold, this approach is able to find the binding site candidates that are undetectable by conventional string match method. When combined with partial sum plot, the hypercomplex cross-correlation is capable of not only detecting the presence of potential binding sites, but also locating the position of the best candidate from the mismatch alignment information of the compared DNA sequence, making consensus sequence matching as accurate as the profiling matching. Furthermore, the hypercomplex cross-correlation is computationally efficient when implemented with the FFT, with computational complexity scaling as $O(N \ln N)$, making this method potentially useful for large computationally intensive tasks, such as database searching.

References

1. Felsenstein J, Sawyer S, Kochin R, An efficient method for matching nucleic-acid sequences, *Nucleic Acids Res* **10**(1):133–139, 1982.
2. Cheever EA, Overton GC, Searls DB, Fast Fourier transform-based correlation of DNA-sequences using complex-plane encoding, *Comput Appl Biosci* **7**(2):143–154, 1991.
3. Peng CK, Buldyrev SV, Goldberger AL, Havlin S, Sciortino F, Simons M, Stanley HE, Long-range correlations in nucleotide-sequences, *Nature* **356**(6365): 168–170, 1992.
4. Anastassiou D, Frequency-domain analysis of biomolecular sequences, *Bioinformatics* **16**(12):1073–1081, 2000.
5. Cristea PD, Conversion of nucleotides sequences into genomic signals, *J Cell Mol Med* **6**(2):279–303, 2002.

6. Brodzik AK, Peters O, Symbol-balanced quaternionic periodicity transform for latent pattern detection in DNA sequences, *Proceedings of the 30th IEEE International Conference on Acoustics, Speech, and Signal Processing* **1–5**:373–376, 2005.
7. Petoukhov SV, Matrix genetics, Part 2: The degeneracy of the genetic code and the octave algebra with two quasi-real units (the genetic octave Yin-Yang-algebra), 1–23, 2008.
8. Rajasekaran S, Jin X, Spouge JL, The efficient computation of position-specific match scores with the fast Fourier transform, *J Comput Biol* **9**(1):23–33, 2002.
9. Cornish-Bowden A, Nomenclature for incompletely specified bases in nucleic-acid sequences — Recommendations 1984, *Nucleic Acids Res* **13**(9):3021–3030, 1985.
10. Shu J-J, Ouw LS, Pairwise alignment of the DNA sequence using hypercomplex number representation, *Bull Math Biol* **66**(5):1423–1438, 2004.
11. Bucher P, Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences, *J Mol Biol* **212**(4):563–578, 1990.
12. Gusfield D, *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology*, Cambridge University Press, 1997.
13. Freschi V, Bogliolo A, Using sequence compression to speedup probabilistic profile matching, *Bioinformatics* **21**(10):2225–2229, 2005.
14. Pei SC, Ding JJ, Chang JH, Efficient implementation of quaternion Fourier transform, convolution, and correlation by 2-D complex FFT, *IEEE Trans Sig Proc* **49**(11):2783–2797, 2001.
15. Bülow T, Sommer G, Hypercomplex signals — A novel extension of the analytic signal to the multidimensional case, *IEEE Trans Sig Proc* **49**(11):2844–2852, 2001.
16. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL, GenBank, *Nucleic Acids Res* **36**(SI):D25–D30, 2008.
17. Prestridge DS, Signal scan — A computer-program that scans DNA-sequences for eukaryotic transcriptional elements, *Comput Appl Biosci* **7**(2):203–206, 1991.