# Graphical Representation for DNA Sequences via Joint Diagonalization of Matrix Pencil

Hong-Jie Yu and De-Shuang Huang, *Senior Member, IEEE*

***Abstract*—Graphical representations provide us with a tool allowing visual inspection of the sequences. To visualize and compare different DNA sequences, a novel alignment-free method is proposed in this paper for both graphical representation and similarity analysis of sequences. We introduce a transformation to represent each DNA sequence with neighboring nucleotide matrix. Then, based on approximate joint diagonalization theory, we transform each DNA primary sequence into a corresponding eigenvalue vector (EVV), which can be considered as numerical characterization of DNA sequence. Meanwhile, we get graphical representation for DNA sequence via the plot of EVV in 2-D plane. Moreover, using *k*-means, we cluster these feature curves of sequences into several reasonable subclasses. In addition, similarity analyses are performed by computing the distances among the obtained vectors. This approach contains more sequence information, and it analyzes all the involved sequence information jointly rather than separately. A typical dendrogram constructed by this method demonstrates the effectiveness of our approach.**

***Index Terms*—Approximate joint diagonalization (AJD), dendrogram, graphical representation, similarity analysis.**

## I. Introduction

GRAPHICAL representations of DNA offer visual inspection of DNA sequences [1]. However, in [2], the author investigated corrections that reveal some aspects of similarity which could not be determined through the traditional alignment-based methods. The space of similarity for complex objects is multidimensional. Complex objects may be similar in one aspect; however, it can be very different in another one. Recently, many numerical characterizations for DNA or protein sequences have been introduced, where most of numerical characterizations are extracted from the string representations and graphical representations. The simpler and more important feature from string representations first used for comparison of genome sequence [3] and later for alignment-free comparison of regulatory sequences [4]. Various frequency-based algorithms later have been introduced for sequence comparisons, as indicated in [5] and [6]. Besides the representations based on single nucleotide, the dinucleotide analysis has also been tried by several authors. Randić [7] proposed a condensed representation of DNA based on pairs of nucleotides. Wu *et al.* [8] proposed the analysis approaches based on neighboring nucleotides of DNA sequence, which reveal the biology information hidden between dual nucleotides.

A recently introduced graphical representation of DNA sequences based on the neighboring dual nucleotides (dinucleotides) [9], [10] is another example of a linear representation. Dual nucleotides can also be divided into groups according to their chemical properties [2] (for review of this topic up to 2011).

DNA sequences can be converted into numerical signals through different transformations. Generally, one can use binary sequences to describe the position of each symbol [11]. The binary representation is certainly one of the earliest and the most popular transformations of DNA. Also, several other different transformation methods have been proposed [12]–[19]. On the transformation methods used in DNA sequences, some do not have a simple numerical interpretation, while others have no biological motivation. Also, some of the representations are irreversible and neglect the sequence structure. So far there has not been an ideal transformation method that is able to analyze every type of correlation among DNA sequences.

In this study, we propose a novel method for graphical representation of DNA sequences and apply the method to sequences analyses. The application of the graphical representation of DNA sequence is illustrated by numerically examining the relationship among different species. The validity of the proposed approach is demonstrated via comparison of correlation coefficients among the results from several methods.

## II. Descriptors of DNA Sequences

Numerical characterization of a graphical representation can also be performed directly from the coordinates or from the properties of the graphs without transforming the graphs to matrices. In this section, we will propose a novel method to transform each DNA sequence into a symmetric sparse matrix, from which feature vector can be extracted finally.

### A. Related Works

Numerical characterizations of both 2-D and 3-D graphical representations for DNA sequences through transforming the

Manuscript received June 20, 2012; revised September 27, 2012; accepted October 22, 2012. Date of publication January 23, 2013; date of current version May 1, 2013. This work was supported in part by the National Science Foundation of China, under Grant 61133010, Grant 31071168, Grant 60905023, Grant 60975005, and Grant 61005010; the special grant of China Postdoctoral Science Foundation (No. 2012T50582); the grant of China Postdoctoral Science Foundation (No. 20100480708); and the Anhui Provincial Natural Science Foundation under 1208085MF96.

H.-J. Yu is with the Department of Mathematics, School of Science, Anhui Science and Technology University, Fengyang 233100, China, and also with the University of Science and Technology of China, Hefei 230027, China (e-mail: yhj70@mail.ustc.edu.cn).

D.-S. Huang is with the Department of Computer Science and Technology, School of Electronics and Information Engineering, Tongji University, Shanghai 200093, China (The Corresponding Author e-mail: huangdeshuang@yahoo.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TITB.2012.2227146

TABLE I
SIXTEEN KINDS OF NEAREST NEIGHBOR NUCLEOTIDE

| $S_j \backslash S_{j+1}$ | A | T | G | C |
|---|---|---|---|---|
| A | AA | AT | AG | AC |
| T | TA | TT | TG | TC |
| G | GA | GT | GG | GC |
| C | CA | CT | CG | CC |

graphs into matrices and deriving the descriptors from these matrices have been widely used by many authors. These descriptors characterizing sequence can be used as the components of similarity measures between a pair of sequences [2]. The examples of similarity analysis of DNA sequences using this method may be found in [15] and [20]–[28].

### B. Construction of Neighboring Matrix for Sequence

Considering a DNA sequence $\boldsymbol{S}$ = "$S_1 S_2 \ldots S_L$," where $S_i \in \{A, T, G, C\}, i = 1, 2, \ldots, L$, and $L$ denotes the length of sequence, there are 16 kinds of dinucleotides in total (as listed in Table I).

Scanning every two adjacent sites successively, such as locus pairs $(S_1, S_2)$, $(S_2, S_3)$, …, $(S_{L-1}, S_L)$, we can obtain a 16 by $(L-1)$ adjacency matrix transformed from the primary sequence via all the adjacent dinucleotides relationship, named as $\boldsymbol{m}$:

$$m = (a_{ij})_{16\times(L-1)}$$

where

$$a_{ij} = \begin{cases} 1, & \text{if } S_j S_{j+1} = \text{the } i\text{th kind of dinucleotides} \\ 0, & \text{otherwise} \end{cases}$$
$$i = 1, 2, \ldots, 16; \quad j = 1, 2, \ldots, L-1.$$

Obviously, the primary biological sequences can be regarded as symbolic signals which may have a rich statistical structure that is the focus of many signal processing algorithms. For example, stochastic symbolic signals are discrete random processes with an unknown amplitude distribution (probability mass function) and a correlation structure [29]. Optimal symbolic-to-digital transformations of the linear, nucleic acid strands into real or complex genomic signals are derived at nucleotide, codon, and amino acid levels. By converting the sequences of nucleotides and polypeptides into digital genomic signals, several approaches offer the possibility to use a large variety of signal processing methods for handling and analyzing the sequences [30]. From the viewpoint of signal processing, the symmetric matrix $\boldsymbol{M}$ can be explained as observations upon 16 kinds of "sensors," i.e., 16 dinucleotides (as depicted in Table I). Thus, the matrix analysis approach in signal processing field can be applied to multiple sequence similarity analysis.

Symmetric matrix has many merits [31]. Using the obtained sparse matrix $\boldsymbol{m}_{16\times(L-1)}$, which is mapped from the primary sequence, we can get a symmetric neighboring matrix $\boldsymbol{M}_{L-1} = \boldsymbol{m}^{\mathrm{T}} * \boldsymbol{m}$ for representing each sequence.

### C. Approximate Joint Diagonalization

Given a matrix pencil, i.e., a set of matrices $\{\boldsymbol{M}^{(1)}, \boldsymbol{M}^{(2)}, \ldots, \boldsymbol{M}^{(N)}\}$, joint diagonalization (JD) refers to the problem of seeking a matrix $\boldsymbol{U}$, which will lead $\boldsymbol{U}^H \boldsymbol{M}^{(n)} \boldsymbol{U}$ to be as diagonal as possible for all $n$, where $\boldsymbol{U}$ is a unitary matrix. In engineering, it is ubiquitous for the use of matrix pencil JD [32], [33]. JD can diagonalize more than two matrices simultaneously. In general, a simple approach to the JD problem is to consider matrix pencil $\{\boldsymbol{M}^{(1)}, \boldsymbol{M}^{(2)}, \ldots, \boldsymbol{M}^{(N)}\}$ consisting of statistical information of the observations that are the estimates of matrices with the form $\boldsymbol{U}^H \boldsymbol{M}^{(n)} \boldsymbol{U}$. When more than two matrices are to be diagonalized, exact diagonalization may also be possible if the matrices possess a certain common structure. Otherwise, one can only speak of approximate joint diagonalization (AJD).

Considerable interest for AJD follows the discovery that it yields a solution for independent component analysis [32] and second-order blind source separation [33], [34].

Meanwhile, from the viewpoint of numerical analysis, the "off" of an $n \times n$ matrix $\boldsymbol{M}$ with entries can be defined as

$$\text{off}(\boldsymbol{M}) \stackrel{\text{def}}{=} \sum_{1\le i\ne j\le n} |m_{ij}|^2 \tag{1}$$

and the unitary diagonalization of a matrix $\boldsymbol{M}$ is equivalent to zeroing **off**$(\boldsymbol{V}^{\mathrm{H}} \boldsymbol{M} \boldsymbol{V})$ by some unitary matrix $\boldsymbol{V}$.

So far, several iterative algorithms have been developed to solve the AJD problem. Generally, for any $n \times n$ matrix $\boldsymbol{V}$, the AJD criterion can be defined as the following nonnegative function of $\boldsymbol{V}$:

$$\boldsymbol{J}(\boldsymbol{V}, \boldsymbol{\Lambda}^{(1)}, \boldsymbol{\Lambda}^{(2)}, \ldots, \boldsymbol{\Lambda}^{(N)}) \stackrel{\text{def}}{=} \sum_{i=1,\ldots,N} \|\boldsymbol{\Lambda}^{(i)} - \boldsymbol{V}^{\mathrm{H}} \boldsymbol{M}^{(i)} \boldsymbol{V}\|^2 \tag{2}$$

that is

$$\boldsymbol{J}(\boldsymbol{V}) \stackrel{\text{def}}{=} \sum_{i=1,\ldots,N} \text{off}\left(\boldsymbol{V}^{\mathrm{H}} \boldsymbol{M}^{(i)} \boldsymbol{V}\right). \tag{3}$$

A unitary matrix is said to be a *joint diagonalizer* of the set $\mathsf{M}$, if it minimizes the AJD criterion over the set of all unitary matrices [see (2)]. First, let us consider the case that each matrix from the set is in the form of $\boldsymbol{M}^{(i)} = \boldsymbol{U} \Lambda^{(i)} \boldsymbol{U}^H$, where $\boldsymbol{\Lambda}^{(i)}$ is a diagonal matrix. Then, obviously $\boldsymbol{J}(\boldsymbol{V}, \boldsymbol{\Lambda}^{(1)}, \boldsymbol{\Lambda}^{(2)}, \ldots, \boldsymbol{\Lambda}^{(N)}) = 0$ holds, and this is the global minimum of the AJD criterion [see (2)] since it holds for any matrix. Thus, if each matrix in the set can be unitarily diagonalized by $\boldsymbol{U}$, obviously, matrix $\boldsymbol{U}$ is a joint diagonalizer of $\mathsf{M}$.

Usually, AJD is *not* required that the involved matrix set $\boldsymbol{M}$ can be exactly simultaneously diagonalized by a common unitary matrix. As a matter of fact, it is *even not* required that the matrices in the set should be *individually* unitarily diagonalizable. This is because we do not require that the "off" function values of all the matrices should be canceled by a unitary transform, and an approximate joint diagonalizer need be just a minimizer of the AJD criterion. Mostly, the AJD criterion cannot be zeroed, and the matrices can only be approximately jointly diagonalized. Hence, an (approximate) joint diagonalizer defines a kind of an "average eigenstructure." This is particularly

convenient for statistical inference where the structural information is to be extracted from sample statistics.

Hence, rather than exactly diagonalizing a single matrix, the AJD allows the information contained in a set of matrices to be integrated in a single unitary matrix [33]. Another important feature of the (possibly approximate) JD is the existence of a numerically efficient algorithm for its computation [31].

### D. Properties of Distance Preserving

In the following, we derive some properties for the proposed AJD-NNM algorithm. Considering two transformations:

1) $\tau_1$: $\mathtt{Sequence}^{(i)} \mapsto \boldsymbol{M}^{(i)}$

$\mathrm{Sequence}^{(i)}$ denotes the $i$th sequence, where the length of sequence is $L$, and $i = 1, 2, \ldots, N$, while $\boldsymbol{M}^{(i)} \in \mathsf{R}^{(L-1)\times(L-1)}$ stands for the corresponding matrices mapped from each primary sequences, and $\boldsymbol{M}^{(i)}$ is a (0, 1) type sparse symmetric matrix, which can be determined by neighboring nucleotide along the $\mathrm{sequence}^{(i)}$.

2) $\tau_2$: $\boldsymbol{M}^{(i)} \mapsto (\lambda_1^{(i)}, \lambda_2^{(i)}, \ldots, \lambda_{L-1}^{(i)})$

The feature vector $\vec{\boldsymbol{F}}_{L-1}^{(i)} = (\lambda_1^{(i)}, \lambda_2^{(i)}, \ldots, \lambda_{L-1}^{(i)})$ is an $(L-1)$-dimensional vector consisting of eigenvalues of $\boldsymbol{M}^{(i)}$ via AJD. So, we can obtain a compound transformation as follows:

$$\tau_2 \circ \tau_1 : \mathtt{Sequence}^{(i)} \mapsto (\lambda_1^{(i)}, \lambda_2^{(i)}, \ldots, \lambda_{L-1}^{(i)}) \tag{4}$$

from which we can freely extract the features of the DNA sequence.

From the viewpoint of algebra space, the transformation can be also presented as

$$\boldsymbol{Ker}\ \boldsymbol{f}: \mathsf{S}^{1\times L} \xrightarrow{\tau} \mathsf{F}^{1\times(L-1)} \tag{5}$$

where $\mathsf{S}^{1\times L}$ denotes the *original sequence space* comprising primary DNA sequence with the length $L$, while $\mathsf{F}^{1\times(L-1)}$ indicates the *objective feature space* that is mapped from the original space. Also, the diagonal elements of $\boldsymbol{\Lambda}$ are just the eigenvalues of the neighboring nucleotide matrix (NNM) via AJD. Furthermore, the obtained data embody the essential property of the primary DNA sequence, as can be seen from the following proposition.

*Definition 1:* The distance $D(s^{(i)}, s^{(j)})$ [35] in the *original sequence space* between two primary sequences, $\mathrm{Sequence}^{(i)}$ and $\mathrm{Sequence}^{(j)}$, is defined as

$$D(s^{(i)}, s^{(j)}) \stackrel{\mathrm{def}}{=} \|\boldsymbol{M}^{(i)} - \boldsymbol{M}^{(j)}\|_{\mathrm{F}} \tag{6}$$

where $\boldsymbol{M}^{(i)}$ is the feature matrix of $\mathrm{sequence}^{(i)}, i, j = 1, 2, \ldots, N$. Here, $\|\boldsymbol{A}\|_{\mathrm{F}} = \sqrt{\mathrm{tr}(\boldsymbol{A}^{\mathrm{H}}\boldsymbol{A})}$ is Frobenius norm of matrix $\mathbf{A}$.

*Definition 2:* Let $\mathbf{R}^{n\times n}$ be a *real normed space* with dimensions $n \times n$, and let $f : \mathbf{R}^{n\times n} \to \mathbf{R}^{1\times n}$ be a function from $\mathbf{R}^{n\times n}$ to $\mathbf{R}^{1\times n}$. A function $f$ is $\alpha$*-distance preserving*, if for any element within the space $\mathbf{R}^{n\times n}$, such as $\boldsymbol{M}^{(i)}$ and $\boldsymbol{M}^{(j)}$, $\|\boldsymbol{M}^{(i)} - \boldsymbol{M}^{(j)}\|_{\mathrm{F}} = \alpha$ implies $\|f(\boldsymbol{M}^{(i)}) - f(\boldsymbol{M}^{(j)})\|_{\mathrm{F}} = \alpha$.

*Theorem 1:* $\tau$: $\mathtt{Sequence}^{(i)} \mapsto (\lambda_1^{(i)}, \lambda_2^{(i)}, \ldots, \lambda_{L-1}^{(i)})$ is a distance-preserving transformation.

*Proof:* Since $\boldsymbol{M}^{(i)}$ and $\boldsymbol{M}^{(j)}$ are the feature matrices of $\mathrm{sequence}^{(i)}$ and $\mathrm{sequence}^{(j)}$, respectively, $i, j = 1, 2, \ldots, N$, let $\lambda(\boldsymbol{M}^{(i)}) = \boldsymbol{V}^{\mathrm{H}}\boldsymbol{M}^{(i)}\boldsymbol{V} = \boldsymbol{\Lambda}^{(i)}$ be a function (see Section II-C); thus, we have

$$\lambda(\boldsymbol{M}^{(i)}) = \mathrm{diag}(\lambda_1^{(i)}, \lambda_2^{(i)}, \ldots, \lambda_{L-1}^{(i)}) \in \mathbf{R}^{(L-1)\times(L-1)}, \quad i = 1, 2, \ldots, N.$$

And hence

$$f(\boldsymbol{M}^{(i)}) = (\lambda_1^{(i)}, \lambda_2^{(i)}, \ldots, \lambda_{L-1}^{(i)}) \stackrel{\mathrm{def}}{=} \vec{\boldsymbol{F}}^{(i)} \in \mathbf{R}^{1\times(L-1)}, \quad i = 1, 2, \ldots, N.$$

By *Definition* 1, we have

$$\begin{aligned}\|\lambda(\boldsymbol{M}^{(i)}) - \lambda(\boldsymbol{M}^{(j)})\|_{\mathrm{F}} &= \|\boldsymbol{V}^{\mathrm{H}}(\boldsymbol{M}^{(i)} - \boldsymbol{M}^{(j)})\boldsymbol{V}\|_{\mathrm{F}} \\ &= \sqrt{\mathrm{tr}[(\boldsymbol{V}^{\mathrm{H}}(\boldsymbol{M}^{(i)} - \boldsymbol{M}^{(j)})\boldsymbol{V})^{\mathrm{H}} * (\boldsymbol{V}^{\mathrm{H}}(\boldsymbol{M}^{(i)} - \boldsymbol{M}^{(j)})\boldsymbol{V})]} \\ &= \sqrt{\mathrm{tr}[\boldsymbol{V}^{\mathrm{H}}(\boldsymbol{M}^{(i)} - \boldsymbol{M}^{(j)})^{\mathrm{H}} * (\boldsymbol{M}^{(i)} - \boldsymbol{M}^{(j)})\boldsymbol{V}]} \\ &= \sqrt{\mathrm{tr}[(\boldsymbol{M}^{(i)} - \boldsymbol{M}^{(j)})^{\mathrm{H}} * (\boldsymbol{M}^{(i)} - \boldsymbol{M}^{(j)})\boldsymbol{V}^{\mathrm{H}}\boldsymbol{V}]} \\ &= \|\boldsymbol{M}^{(i)} - \boldsymbol{M}^{(j)}\|_{\mathrm{F}} = \alpha.\end{aligned}$$

By *Definition* 2, we know that the following equation hold:

$$\|f(\boldsymbol{M}^{(i)}) - f(\boldsymbol{M}^{(j)})\|_{\mathrm{F}} = \|\lambda(\boldsymbol{M}^{(i)}) - \lambda(\boldsymbol{M}^{(j)})\|_{\mathrm{F}} = \alpha.$$

For a given primary sequence, $\mathrm{s}_L^{(i)}$, there exists a unique $(L-1)$-dimensional eigenvalue vector (EVV) $(\lambda_1^{(i)}, \lambda_2^{(i)}, \ldots, \lambda_{L-1}^{(i)}) = \vec{\boldsymbol{F}}_L^{(i)}$ obtained via the proposed approach of AJD-NNM. That is, $\vec{\boldsymbol{F}}_L^{(i)}$ depends only on the primary sequence and its length, which can be written as

$$f(\mathrm{s}_L^{(i)}) = \vec{\boldsymbol{F}}_L^{(i)}, \quad i = 1, 2, \ldots, N.$$

Here, superscript $i$ denotes the label of sequence, and $L$ indicates the length of sequence.

By *Definition* 2, it can be seen that the ensemble transformation $\tau$ is indeed a distance-preserving transformation. $\boldsymbol{QED}$

Based on *Definition* 2 and *Theorem* 1, we can calculate all the EVVs of each obtained NNM, such as $\vec{\boldsymbol{F}}_{L-1}^{(i)} = (\lambda_1^{(i)}, \lambda_2^{(i)}, \ldots, \lambda_{L-1}^{(i)}), i = 1, 2, \ldots, N$, where $L$ denotes the common least length of $N$ primary DNA or protein sequences. Then, we can get $N$ corresponding $(L-1)$-dimensional vectors, which can be regarded as features extracted from the original DNA sequence. The steps of AJD-NNM algorithm can be summarized as follows in the top of the next page.

## III. Graphical Representations

Alternative to the time-consuming alignment methods, graphical representation approaches reveal different aspects of similarity, offering numerical characterizations of both similarity and visualization. However, almost all these methods can only represent each sequence *separately* rather than *jointly*. In this section, a novel method of graphical representation for DNA sequences is presented, which can *jointly* consider the mutual information among all the involved sequences.

We selected the dataset that comprises 11 sequences of the first exon in the *beta*-globin gene, shown in Table II.

**Input**: multiple biological sequences with truncated common length $L$: $\boldsymbol{S}^{(1)}, \boldsymbol{S}^{(2)}, \ldots, \boldsymbol{S}^{(N)}$
**Initialize**: *Tol* - An imposed *tolerance* on the change in objective function for a stopping condition
**begin**
  **for** $n$=1 **to** $N$ **do**
    Transform original sequences $\boldsymbol{S}^{(n)}$ into ($L$-1) by ($L$-1) sparse symmetric matrix $\boldsymbol{M}^{(n)}$
  **end for**
  Consider the obtained matrix set $\mathsf{M} = \{\boldsymbol{M}^{(1)}, \boldsymbol{M}^{(2)}, \cdots, \boldsymbol{M}^{(N)}\}$ and objective function
  $J(\boldsymbol{V}, \boldsymbol{\Lambda}^{(1)}, \boldsymbol{\Lambda}^{(2)}, \cdots, \boldsymbol{\Lambda}^{(N)}) = \sum_{i=1,\cdots,N} \left\| \boldsymbol{\Lambda}^{(i)} - \boldsymbol{V}^{\mathrm{H}} \boldsymbol{M}^{(i)} \boldsymbol{V} \right\|^2$
  **while** $J(\boldsymbol{V}, \boldsymbol{\Lambda}^{(1)}, \boldsymbol{\Lambda}^{(2)}, \cdots, \boldsymbol{\Lambda}^{(N)}) \geq tolerance$ **do**
    Update $\boldsymbol{V}$ using AC-DC algorithm [36]
  **end while**
  **for** $n$=1 **to** $N$ **do**
    $\boldsymbol{F}^{(n)} \leftarrow \mathrm{diag}(\boldsymbol{\Lambda}^{(n)})$
    Plot and categorize the $N$ feature curves with $\boldsymbol{F}^{(n)}$
  **end for**
  **for** $n$=1 **to** $N$-1 **do**
    **for** $m$=$n$+1 **to** $N$ **do**
      Calculate pairwise distances using $\boldsymbol{F}^{(n)}$ by
      $\mathrm{D}(\mathrm{s}^{(i)}, \mathrm{s}^{(j)}) = \left\| \boldsymbol{M}^{(i)} - \boldsymbol{M}^{(j)} \right\|_{\mathrm{F}} = \left\| \bar{\boldsymbol{F}}^{(i)} - \bar{\boldsymbol{F}}^{(j)} \right\|_{\mathrm{F}}$
    **end for**
  **end for**
  Draw the dendrogram using the pairwise distances matrix
**end**

### A. Calculation of EVVs

In order to make multisequences be comparable, we truncated each sequence from the 1st site to 86th site, since the common least length of those eleven sequences is 86. According to the procedure for AJD-NNM algorithm depicted in Section II-D, all the 85-tuple EVVs are calculated via AJD upon all the 11 NNMs. These vectors were orderly connected head and tail using the walk strategy. Thus, there are 11 curves plotted, as shown in Fig. 1.

### B. Convergence Analysis of the AJD Algorithm

Based on the Frobenius-norm formulation, Ziehe *et al.* [36] investigated a fast algorithm for joint diagonalization problem and provided a comparison to other leading diagonalization methods, such as the extended Jacobi method as used in the JADE algorithm of Cardoso and Souloumiac [32] (orthogonal Frobenius norm formulation), Pham's algorithm for positive-definite matrices [34] and Yeredor's AC–DC algorithm [37] (nonorthogonal, subspace fitting formulation).

On the convergence analysis of AJD, there are two criteria, i.e., the cost function and the convergence ratio. Based on the cost function criterion, Yeredor [37] proposed an iterative algorithm (AC–DC) for AJD of a given set of matrices in the weighted least square (LS) sense with arbitrary positive

TABLE II
CONCISE INFORMATION OF *BETA*-GLOBIN GENE SEQUENCES

| Species | Coding Sequences |
|---|---|
| Human | ATGGTGCATCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGG |
| (92 bases) | GCAAGGTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAG |
| Goat | ATGCTGACTGCTGAGGAGAAGGCTGCCGTCACCGGCTTCTGGGGCAAG |
| (86 bases) | GTGAAAGTGGATGAAGTTGGTGCTGAGGCCCTGGGCAG |
| Opossum | ATGGTGCACTTGACTTCTGAGGAGAAGAACTGCATCACTACCATCTGG |
| (92bases) | TCTAAGGTGCAGGTTGACCAGACTGGTGGTGAGGCCCTTGGCAG |
| Gallus | ATGGTGCACTGGACTGCTGAGGAGAAGCAGCTCATCACCGGCCTCTGG |
| (92 bases) | GGCAAGGTCAATGTGGCCGAATGTGGGGCCGAAGCCCTGGCCAG |
| Lemur | ATGACTTTGCTGAGTGCTGAGGAGAATGCTCATGTCACCTCTCTGTGGG |
| (92 bases) | GCAAGGTGGATGTAGAGAAAGTTGGTGGCGAGGCCTTGGGCAG |
| Mouse | ATGGTGCACCTGACTGATGCTGAGAAGGCTGCTGTCTCTTGCCTGTGGG |
| (92 bases) | GAAAGGTGAACTCCGATGAAGTTGGTGGTGAGGCCCTGGGCAG |
| Rabbit | ATGGTGCATCTGTCCAGTGAGGAGAAGTCTGCGGTCACTGCCCTGTGGG |
| (92 bases) | GCAAGGTGAATGTGGAAGAAGTTGGTGGTGAGGCCCTGGGCAG |
| Rat | ATGGTGCACCTAACTGATGCTGAGAAGGCTACTGTTAGTGGCCTGTGGG |
| (92 bases) | GAAAGGTGAACCCTGATAATGTTGGCGCTGAGGCCCTGGGCAG |
| Gorilla | ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGG |
| (93 bases) | GCAAGGTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAGG |
| Bovine | ATGCTGACTGCTGAGGAGAAGGCTGCCGTCACCGCCTTTTGGGGCAAG |
| (86 bases) | GTGAAAGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAG |
| Chimpanzee | ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAG |
| (105 bases) | GTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAGGTTGGTATCAAGG |

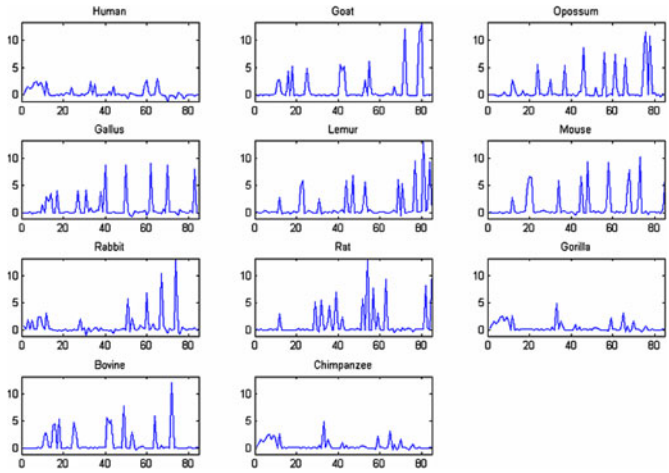


Fig. 1. Graphical representation of the first exon in the *beta*-globin gene from eleven species based on 85-tuple EVVs via AJD upon all 11 NNMs. The $y$-axis indicates the values of each element in feature vectors $(\lambda_1, \lambda_2, \ldots, \lambda_{L-1})$.

weights. Convergence to a stationary point of the LS criterion is guaranteed under mild conditions. The author has proved the convergence of AC–DC algorithm in theory [37].

The diagonalizer matrix $\boldsymbol{V}$ of the objective function at the optimum point depends on the least error, which is of course unknown. Here, we selected the AC–DC optimization scheme, for it can still work regardless of initialization. In the following, let us consider a simplified expression of the least error:

$$\mathrm{Err}(j) \stackrel{\mathrm{def}}{=} \sum_{i=1,\cdots,N} \| \boldsymbol{\Lambda}_j^{(i)} - \boldsymbol{V}_j^{\mathrm{H}} \boldsymbol{M}^{(i)} \boldsymbol{V}_j \|_{\mathrm{F}}^2 \qquad (7)$$

where $j$ ranges from 2 to the largest number of iterations.

Let $\varepsilon$ be the predefined error threshold. If $|\mathrm{Err}(j+1) - \mathrm{Err}(j)| < \varepsilon$, i.e., the error does not change strongly from the $j$th iteration to the $(j+1)$th one, then either the obtained diagonalizer $\boldsymbol{V}$ or the diagonalized $\boldsymbol{\Lambda}^{(i)}$ is optimal just at the $j$th step.

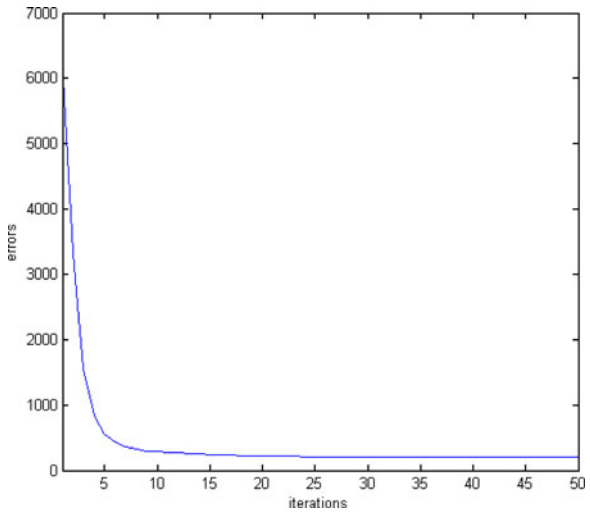


Fig. 2. Curve for the errors of the proposed algorithm (at AJD stage) versus iterations.

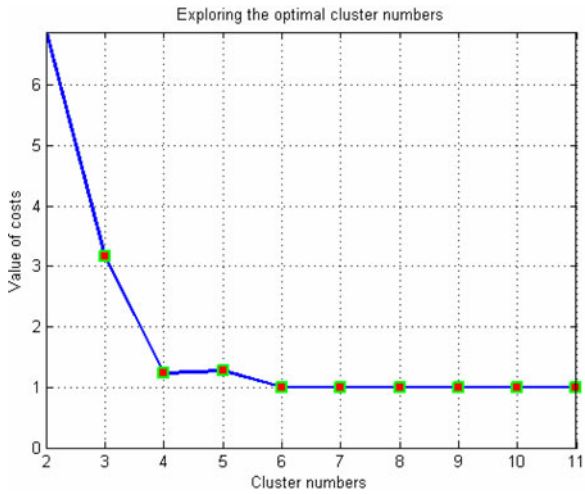


Fig. 3. Curve for the costs of $k$-means clustering for the 11 sequences to explore the optimal cluster numbers. The $x$-axis indicates the predefined cluster numbers, and the $y$-axis denotes the corresponding costs based on squared error. There is a minimum point at $k = 4$, which suggests that four should be the number of clusters.

This enables us to draw the conclusions about the convergence behavior of gradient-based optimization algorithms. The experimental results are shown in Fig. 2, from which we can find that the AJD algorithm can fast converge just at the fifth iteration.

### C. Clustering of Sequences Based on Their EVVs

The most well-known hierarchical algorithms are single-link and complete-link; the most popular and the simplest partitional algorithm is $k$-means. Even though $k$-means was first proposed over 50 years ago, it is still one of the most widely used algorithms for clustering. Ease of implementation, simplicity, efficiency, and empirical success are the main reasons for its popularity.

$K$-means starts with an initial partition with $k$ clusters and assign patterns to clusters so as to reduce the squared error. Since the squared error always decreases with an increase in the number of clusters $k$, it can be minimized only for a fixed number of clusters.

The $k$-means algorithm requires three user-specified parameters: number of clusters $k$, cluster initialization, and distance metric. The most critical choice is $k$. While no perfect mathematical criterion exists, a number of heuristics are available for choosing $k$. Typically, $k$-means is run independently for different values of $k$ and the partition that appears the most meaningful to the domain expert is selected. Different initializations can lead to different final clustering because $k$-means only converges to local minima. One way to overcome the local minima is to run the $k$-means algorithm, for a given $k$, with multiple different initial partitions and choose the partition with the smallest squared error.

In this study, the dimension of EVVs is high, which is up to $L - 1$, i.e., 85, in this dataset. As a result, $k$-means is used with the "Correlation" metric for computing the distance between points and cluster centers. Fig. 3 shows the changes of value of costs with respect to $k$. When $k$ is changed from 3 to 5, there is a significant drop of the cost value at $k = 4$, which is also a minimum point. This suggests that four subclasses should be reasonable for this dataset. Thus, we can cluster these 11 feature curves into four subclasses via $k$-means. From Fig. 4, it can be seen that

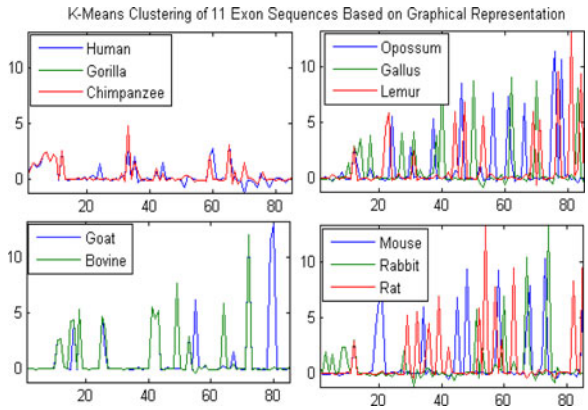


Fig. 4. Four clusters via $k$-means clustering for the 11 sequences according to their curves shown in Fig. 1. The optimal cluster numbers is four, which is explored by the costs index shown in Fig. 3.

1) Human, Gorilla, and Chimpanzee are close to each other;
2) Goat is the closest to Bovine, for their curves are almost coinciding with each other;
3) the group (Opossum, Gallus and Lemur) is far away from the other three groups;
4) Mouse, Rabbit, and Rat fall into the last category.

In a word, these phenomena are also in accordance with the evolution fact from the viewpoint of evolutionary relationships of organisms.

## IV. Similarity Analysis

The comparison based on sequence descriptors is another method, which has been routinely used in similarity analysis.

Here, we use the aforementioned 11 EVVs for quantitatively comparing different DNA sequences.

### A. Calculation of Pairwise Distances

On the other hand, we applied the numerical characterization to examine the similarity of sequences in the dataset. Just as depicted in Section II-D, as usual, we selected the "Euclidean" metric to calculate the genetic distance in this study. By Theorem 1, we need only to calculate the Euclidean distance between every two EVVs: $\vec{\boldsymbol{F}}^{(i)}$ and $\vec{\boldsymbol{F}}^{(j)}$, which are obtained by the aforementioned approach AJD-NNM. According to Definitions 1 and 2, it can be found that the dissimilarity degree can be determined through Euclidean distance between the every two sequences (listed in Table I). The Euclidean distance between the $i$th and the $j$th sequences can be calculated as

$$D(\boldsymbol{S}^{(i)}, \boldsymbol{S}^{(j)}) \stackrel{\text{def}}{=} \|\boldsymbol{M}^{(i)} - \boldsymbol{M}^{(j)}\|_{\text{F}} = \|\vec{\boldsymbol{F}}^{(i)} - \vec{\boldsymbol{F}}^{(j)}\|_{\text{F}} \quad (8)$$

where $\vec{\boldsymbol{F}}^{(i)} = (\lambda_1^{(i)}, \lambda_2^{(i)}, \ldots, \lambda_{L-1}^{(i)})$ denotes the feature vectors through distance-preserving transformation from the primary sequence with the length $L = 86$, while $\|\bullet\|_{\text{F}}$ indicates the *Frobenius* norm of a matrix or vector. Obviously, the smaller distance means the two sequences are more similar. The data for the comparison of the similarity among the 11 coding sequences were obtained by calculating the Euclidean distance $D(\boldsymbol{S}^{(i)},\boldsymbol{S}^{(j)})$.

### B. Phylogeny of 11 Beta-Globin Genes

The alphabet representation of biological sequences is easily handled with computer but difficult for us to observe their differences [38]. Phylogenetic tree provides us with a simple way to view various biological sequences and facilitates sequence comparison with the intuitive pictures and pattern. The proposed approach, i.e., AJD-NNM, was further tested on phylogenetic analysis. Given a set of biological sequences, their phylogenetic relationship can be obtained through the following main operations.

1) First, we calculated the $(L-1)$-dimensional EVV of each biological sequence through AJD-NNM.
2) Second, we got the similarity distance with the Euclidean metric.
3) Third, by arranging all the similarity distance into a matrix, we obtained a pairwise distance matrix.
4) Finally, based on the pairwise distance matrix, we plotted the dendrogram with MATLAB code.

The experimental results are given in Fig. 5 and Table III. From Fig. 5, we can find that the 11 species are separated clearly.

### C. Comparison With Representative Works

Generally, the validation of a newly proposed alignment-free algorithm can be implemented through the comparison with the traditional alignment-based approach. We calculated the pairwise distances of these 11 sequences using MEGA software based on alignment framework, so that we can make a comparison analysis subsequently. The evolutionary history was inferred using the neighbor-joining method [39]. The alignment-based results of the pairwise distances are listed in Table IV.

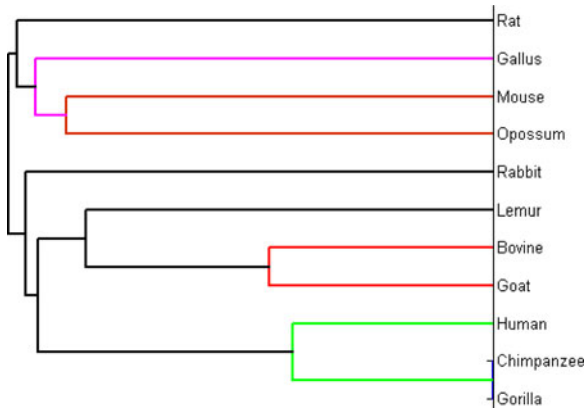


Fig. 5. Dendrogram for the 11 sequences according to the pairwise distance listed in Table III.

To compare the results from our approach (see Table III) with alignment-based approach (shown in Table IV) and other related works, we listed the recently published results in Table V on comparing the similarity among Human and other several representative species, where each entry was extracted from the first row (or column) in the specific tables from the corresponding bibliography. Table V shows that the six selected representative species can be categorized into three groups according to the distances between itself and Human.

1) The closest one belonging to Human is the first group: Chimpanzee and Gorilla (see the third and fourth columns in Table V).
2) Next group includes the species (see columns 5 and 6 of Table V).
3) The last group (Opossum and Gallus) is far away from Human in the light of evolutionary relationship (see the last two columns from Table V).

There is an overall agreement among the similarities obtained by different approaches despite some variation among them. In addition, the results obtained from most of these methods are also consistent with the evolutionary fact.

For the purpose of intuitive analysis, we calculated the correlation coefficients between the results via alignment-based approach (the bottom of Table V) and each of the rest 17 results (the rest rows of Table V), respectively. The experimental results of correlation degree are given in Fig. 6. As can be seen from Fig. 6, our result (No. 17) has the highest correlation degree with that obtained by the traditional alignment-based method using MEGA 4. However, there are two extreme cases among the other 16 different representative methods, such as the tenth and the fourteenth cases.

*Case 1:* As for the 14th work [40], the correlation coefficient value is –0.048. The disagreements are appearing in their results, where the distances have remarkable difference between the similar species pairs such as Human–Chimpanzee and Human–Gorilla. The distance of Human–Gorilla is 0.0424, while the distance of Human–Chimpanzee is 0.0062. The former is about 6.8387 times the latter, which are not consistent with the

TABLE III
PAIRWISE DISTANCE MATRIX OF ELEVEN EXON 1 FROM *BETA*-GLOBIN GENE SEQUENCES WITH OUR MODEL VIA "EUCLIDEAN" METRIC

| Species $i\backslash j$ | Goat | Opossum | Gallus | Lemur | Mouse | Rabbit | Rat | Gorilla | Bovine | Chimpanzee |
|---|---|---|---|---|---|---|---|---|---|---|
| Human | 26.8209 | 25.9952 | 23.5869 | 25.2515 | 25.8007 | 20.5706 | 27.0102 | 5.3704 | 22.4257 | 5.3704 |
| Goat | | 35.6829 | 33.5008 | 34.6771 | 34.7092 | 31.7451 | 36.2950 | 26.7804 | 21.4802 | 26.7804 |
| Opossum | | | 32.6023 | 34.8208 | 34.8006 | 32.4892 | 36.4528 | 26.3341 | 32.7734 | 26.3341 |
| Gallus | | | | 32.1543 | 32.3047 | 29.7515 | 34.0423 | 22.3371 | 29.9851 | 22.3371 |
| Lemur | | | | | 34.3885 | 30.9739 | 35.8069 | 25.7036 | 31.4534 | 25.7036 |
| Mouse | | | | | | 29.3939 | 33.3897 | 25.8560 | 31.5484 | 25.8560 |
| Rabbit | | | | | | | 32.4424 | 20.8903 | 28.4722 | 20.8903 |
| Rat | | | | | | | | 27.2888 | 33.2117 | 27.2888 |
| Gorilla | | | | | | | | | 22.3227 | 0.0000 |
| Bovine | | | | | | | | | | 22.3227 |

TABLE IV
PAIRWISE DISTANCE (NEIGHBOR JOINING) MATRIX OF11 EXON 1 SEQUENCES VIA ALIGNMENT-BASED APPROACH THROUGH MEGA 4

| Species $i\backslash j$ | Goat | Opossum | Gallus | Lemur | Mouse | Rabbit | Rat | Gorilla | Bovine | Chimpanzee |
|---|---|---|---|---|---|---|---|---|---|---|
| Human | 0.6628 | 0.2674 | 0.2791 | 0.2791 | 0.6977 | 0.1163 | 0.2093 | 0.0116 | 0.6512 | 0.0116 |
| Goat | | 0.6744 | 0.6163 | 0.6163 | 0.6163 | 0.6395 | 0.6279 | 0.6512 | 0.0349 | 0.6512 |
| Opossum | | | 0.3023 | 0.4070 | 0.7326 | 0.3140 | 0.3837 | 0.2674 | 0.6744 | 0.2674 |
| Gallus | | | | 0.3721 | 0.6977 | 0.3023 | 0.3721 | 0.2674 | 0.6163 | 0.2674 |
| Lemur | | | | | 0.7093 | 0.2907 | 0.3721 | 0.2791 | 0.6279 | 0.2791 |
| Mouse | | | | | | 0.6977 | 0.7209 | 0.7093 | 0.6279 | 0.7093 |
| Rabbit | | | | | | | 0.2791 | 0.1047 | 0.6279 | 0.1047 |
| Rat | | | | | | | | 0.2093 | 0.6163 | 0.2093 |
| Gorilla | | | | | | | | | 0.6395 | 0.0000 |
| Bovine | | | | | | | | | | 0.6395 |

TABLE V
RESULTS COMPARISON WITH REPRESENTATIVE WORKS UPON THE SIMILARITY DEGREE HUMAN VERSUS OTHER SEVERAL SPECIES

| No. | Results from Ref. | Gorilla | Chimpanzee | Bovine | Goat | Gallus | Opossum |
|---|---|---|---|---|---|---|---|
| 1 | Randić et al.(2003), Table 3 [21] | 0.021 | 0.017 | 0.084 | 0.061 | 0.109 | 0.148 |
| 2 | Liao and Wang(2004),Table 7 [28] | 0.054547 | 0.065209 | 0.096158 | 0.108362 | 0.060439 | 0.130445 |
| 3 | Liao et al.(2005),Table 4 [42] | 0.022047 | 0.020569 | 0.040787 | 0.043579 | 0.088281 | 0.079935 |
| 4 | Chi and Ding(2005), Table 6 [43] | 0.0012 | 0.0094 | 0.0580 | 0.0162 | 0.0133 | 0.0601 |
| 5 | Liao and Ding(2006),Table 4 [44] | 0.044067 | 0.040705 | 0.081376 | 0.086956 | 0.176581 | 0.159761 |
| 6 | Liu et al.(2006), Table 5 [45] | 0.0079 | 0.0145 | 0.0750 | 0.1078 | 0.2417 | 0.2815 |
| 7 | Zhang and Chen(2006), Table 6 [46] | 0.0133 | 0.0093 | 0.0455 | 0.0413 | 0.0595 | 0.0455 |
| 8 | Yao et al.(2008), Table 5 [47] | 0.00162 | 0.00143 | 0.00319 | 0.00556 | 0.00721 | 0.00547 |
| 9 | Wang et al.(2009) Table 3 [16] | 0.0146 | 0.0131 | 0.0259 | 0.0326 | 0.0870 | 0.0813 |
| 10 | Zhang (2009), Table 1 [41] | 263.3 | ***957.2*** | 360.6 | 476.9 | 1155.9 | 1186.3 |
| 11 | Tang et al.(2010), Table 3 [48] | 0.0441 | 0.0399 | 0.0799 | 0.0869 | 0.1766 | 0.1598 |
| 12 | Luo et al. (2010), Table III [49] | 0.0136 | 0.0152 | 0.0372 | 0.0451 | 0.0969 | 0.0962 |
| 13 | Luo et al. (2010), Table IV [49] | 0.0115 | 0.0129 | 0.0555 | 0.0970 | 0.1466 | 0.1050 |
| 14 | Xie et al.(2011), Table 3 [40] | ***0.0424*** | 0.0062 | 0.0735 | 0.0789 | 1.1475 | 0.6468 |
| 15 | Li et al.(2010), Table 7 [50] | 0.010 | 0.011 | 0.100 | 0.105 | 0.089 | 0.215 |
| 16 | Bielińska (2011), Table 4 [2] | 46.50 | 46.50 | 46.55 | 46.80 | 45.69 | 46.22 |
| 17 | **Table III in [This work]** | **5.3704** | **5.3704** | **22.4257** | **26.8209** | **23.5869** | **25.9952** |
| | **Table IV in [This work]** | **0.0116** | **0.0116** | **0.6512** | **0.6628** | **0.2791** | **0.2674** |

evolutional facts. In addition, the results from most works support the conclusion that both the distance of Human–Chimpanzee and that of Human–Gorilla are far lower than the others, because Human, Chimpanzee, and Gorilla are relatively very close to each other.

*Case 2:* Concerning the 10th work [41], the correlation coefficient is the lowest (–0.2933). Similarly, the inconsistency is also appearing in their results, where the Euclidean distance of the species pairs of Human–Chimpanzee (957.2) is far greater than that of Human–Gorilla (263.3), and is even far greater than those of Human–Goat, Human–Mouse, and Human–Bovine.

Compared with the two related works mentioned previously [40], [41], the results derived from our method are also closer to the evolutional facts, as we have known.

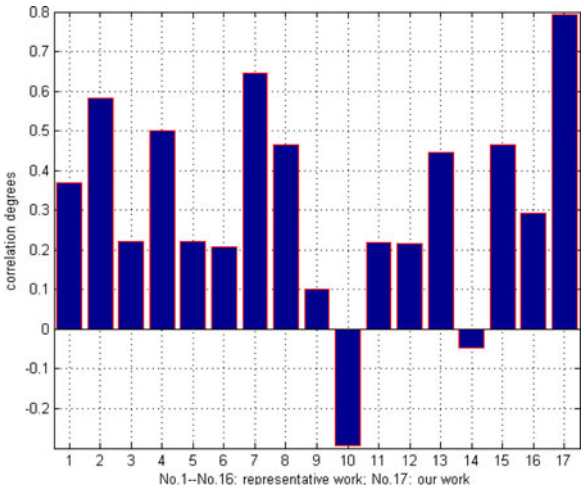


Fig. 6. Comparison of similarity degree with the alignment-based approach via MEGA. The $x$-axis denotes the related works (No.1–No.16) and our work (No. 17), respectively, while the $y$-axis indicates the correlation degrees (correlation coefficients values) of the pairwise distances between Human and other several species in Table II.

## V. Conclusions

The proposed approach (AJD-NNM) has two stages: NNM and AJD. NNM grasps the sequential property of biological sequence, where the sequential property was considered at the stage of transforming sequences into symmetric matrices. Another obvious advantage of our approach over others is that the sequence comparison is based on the information lossless technique. Therefore, it has great improvement on the precision of similarity analysis, which is illustrated through Fig. 6.

Particularly, at the second stage, AJD extracted the features from multiple sequences *jointly* rather than *separately*, which can simultaneously discover that some subgroups of organism have common structure at molecular level while others have not. In addition, we investigated the optimal cluster numbers according to the changes of squared error with the increase of cluster numbers, which is served as criterion. Thus, the results are more objective rather than subjective. Then, we grouped the curves into four subclasses via $k$-means. The grouped results are consistent with the evolutionary fact, which demonstrate that the proposed graphical representation is reasonable. Finally, it is worth noting that our distance measures do not use any evolutionary model, and our approach has the fine property of distance preserving (see Theorem 1).

Therefore, based on the AJD from the NNM, it can be seen that the proposed approach provides us with a reasonable way to compare different biological sequences. In the future, we are planning to enhance our algorithm to apply upon the dataset with a longer sequence length from DNA or protein sequence.

## Acknowledgment

The authors would like to thank the anonymous reviewers for many valuable suggestions that have improved this manuscript.

## References

[1] A. Nandy, M. Harle, and S. C. Basak, "Mathematical descriptors of DNA sequences: Development and applications," *ARKIVOC*, vol. ix, pp. 211–238, 2006.
[2] D. Bielińska-Wż, "Graphical and numerical representations of DNA sequences: Statistical aspects of similarity," *J. Math. Chem.*, vol. 49, pp. 2345–2407, 2011.
[3] B. E. Blaisdell, "A measure of the similarity of sets of sequences not requiring sequence alignment," *Proc. Natl. Acad. Sci.*, vol. 83, pp. 5155–5159, 1986.
[4] M. R. Kantorovitz, G. E. Robinson, and S. Sinha, "A statistical method for alignment-free comparison of regulatory sequences," *Bioinformatics*, vol. 23, pp. i249–i255, 2007.
[5] G. E. Sims, S. R. Jun, G. A. Wu, and S. H. Kim, "Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions," *Proc. Natl. Acad. Sci.*, vol. 106, pp. 2677–2682, 2009.
[6] S. R. Jun, G. E. Sims, G. A. Wu, and S. H. Kim, "Whole-proteome phylogeny of prokaryotes by feature frequency profiles: An alignment-free method with optimal feature resolution," *Proc. Natl. Acad. Sci.*, vol. 107, pp. 133–138, 2009.
[7] M. Randić, "Condensed representation of DNA primary sequences," *J. Chem. Inf. Comput. Sci.*, vol. 40, pp. 50–56, 2000.
[8] Y. Wu, A. W.-C. Liew, H. Yan, and M. Yang, "DB-Curve: A novel 2D method of DNA sequence visualization and representation," *Chem. Phys. Lett.*, vol. 367, pp. 170–176, 2003.
[9] Z. B. Liu, B. Liao, and W. Zhu, "A new method to analyze the similarity based on dual nucleotides of the DNA sequence," *MATCH*, vol. 61, pp. 541–552, 2009.
[10] Z. Liu, B. Liao, W. Zhu, and G. Huang, "A 2D graphical representation of DNA sequence based on dual nucleotides and its application," *Int. J. Quantum Chem.*, vol. 109, pp. 948–958, 2009.
[11] R. F. Voss, "Evolution of long-rang fractal correlations and 1/f noise in DNAbase sequences," *Phys. Rev. Lett.*, vol. 68, pp. 3805–3808, 1992.
[12] M. Akhtar, J. Epps, and E. Ambikairajah, "On DNA numerical representation for period-3 based exon prediction," in *Proc. 5th Int. Workshop Genomic Signal Process. Stat.*, Tuusula, Finland, 2007, pp. 1–4.
[13] H. J. Jeffrey, "Chaos game representation of gene structure," *Nucleic Acids Res.*, vol. 18, pp. 2163–2170, 1990.
[14] R. Zhang and C. T. Zhang, "Z curves, an intutive tool for visualizing and analyzing the DNA sequences," *J. Biomol. Struct. Dyn.*, vol. 11, pp. 767–782, 1994.
[15] M. Randić, "Another look at the chaos-game representation of DNA," *Chem. Phys. Lett.*, vol. 456, pp. 84–88, 2008.
[16] S. Wang, F. Tian, W. Feng, and X. Liu, "Applications of representation method for DNA sequences based on symbolic dynamics," *J. Mol. Struct. THEOCHEM*, vol. 909, pp. 33–42, 2009.
[17] A. K. Brodzik and O. Peters, "Symbol-balanced quaternionic periodicity transform for latent pattern detection in DNA sequences," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Philadelphia, PA, 2005, pp. 373–376.
[18] P. D. Cristea, "Large scale features in DNA genomic signals," *Signal Process.*, vol. 83, pp. 871–888, 2003.
[19] B. Liao, M. Tan, and K. Ding, "Application of 2-D graphical representation of DNA sequence," *Chem. Phys. Lett.*, vol. 414, pp. 296–300, 2005.
[20] M. Randić, M. Vracko, N. Lers, and D. Plavsic, "Novel 2-D graphical representation of DNA sequences and their numerical characterization," *Chem. Phys. Lett.*, vol. 368, pp. 1–6, 2003.
[21] M. Randić, M. Vracko, N. Lers, and D. Plavsic, "Analysis of similarity/dissimilarity of DNA sequences based on a novel 2-D graphical representation," *Chem. Phys. Lett.*, vol. 371, pp. 202–207, 2003.
[22] J. Song and H. W. Tang, "A new 2-D graphical representation of DNA sequences and their numerical characterization," *J. Biochem. Biophys. Methods*, vol. 63, pp. 228–239, 2005.
[23] B. Liao and T. M. Wang, "New 2D graphical representations of DNA sequences," *J. Comput. Chem.*, vol. 25, pp. 1364–1368, 2004.
[24] Y.-H. Yao and T.-M. Wang, "A class of new 2-D graphical representation of DNA sequences and their application," *Chem. Phys. Lett.*, vol. 398, pp. 318–323, 2004.
[25] M. Randić, M. Vračko, J. Zupan, and M. Novič, "Compact 2-D graphical representation of DNA," *Chem. Phys. Lett.*, vol. 373, pp. 558–562, 2003.
[26] M. Randić, "Graphical representations of DNA as 2-D map," *Chem. Phys. Lett.*, vol. 386, pp. 468–471, 2004.
[27] L. Yang, X. Zhang, and H. Zhu, "Alignment free comparison: Similarity distribution between the DNA primary sequences based on the shortest absent word," *J. Theor. Biol.*, vol. 295, pp. 125–131, 2012.

[28] B. Liao and T.-M. Wang, "Analysis of similarity/dissimilarity of DNA sequences based on 3-D graphical representation," *Chem. Phys. Lett.*, vol. 388, pp. 195–200, 2004.
[29] W. Wang and D. H. Johnson, "Computing linear transforms of symbolic signals," *IEEE Trans. Signal Process.*, vol. 50, no. 3, pp. 628–634, Mar. 2002.
[30] P. D. Cristea, "Conversion of nucleotide sequences into genomic signals," *J. Cell Mol. Med.*, vol. 6, pp. 279–303, 2002.
[31] G. H. Golub and C. F. V. Loan, *Matrix Computations*, 3rd ed. Baltimore, MD: Johns Hopkins Univ. Press, 1996.
[32] J.-F. Cardoso and A. Souloumiac, "Blind beamforming for non-Gaussian signals," *IEE Proc-F (Radar Signal Process)*, vol. 140, pp. 362–370, 1993.
[33] A. Belouchrani, K. Abed-Meraim, J.-F. Cardoso, and R. Moulines, "A blind source separation technique using second-order statistics," *IEEE Trans. Signal Process.*, vol. 45, no. 2, pp. 434–444, Feb. 1997.
[34] D.-T. Pham and J.-F. Cardoso, "Blind separation of instantaneous mixtures of non stationary sources," *IEEE Trans. Signal Process.*, vol. 42, no. 9, pp. 1837–1848, Sep. 2001.
[35] J. R. Magnus and H. Neudecker, *Matrix Differential Calculus with Applications in Statistics and Econometrics*. New York: Wiley, 1999.
[36] A. Ziehe, P. Laskov, G. Nolte, and K.-R. Muller, "A fast algorithm for joint diagonalization with non-orthogonal transformations and its application to blind source separation," *J. Mach. Learn. Res.*, vol. 5, pp. 777–800, 2004.
[37] A. Yeredor, "Non-orthogonal joint diagonalization in the least-squares sense with application in blind source separation," *IEEE Trans. Signal Process.*, vol. 50, no. 7, pp. 1545–1553, Jul. 2002.
[38] Q. Dai, X. Guo, and L. Li, "Sequence comparison via polar coordinates representation and curve tree," *J. Theor. Biol.*, pp. 78–85, 2011.
[39] N. Saitou and M. Nei, "The neighbor-joining method: A new method for reconstructing phylogenetic trees," *Mol. Biol. Evol.*, vol. 4, pp. 406–425, 1987.
[40] G. Xie and Z. Mo, "Three 3D graphical representations of DNA primary sequences based on the classifications of DNA bases and their applications," *J. Theor. Biol.*, vol. 269, pp. 123–130, 2011.
[41] Z. J. Zhang, "DV-Curve: a novel intuitive tool for visualizing and analyzing DNA sequences," *Bioinformatics*, vol. 25, pp. 1112–1117, 2009.
[42] B. Liao, M. Tan, and K. Ding, "A 4D representation of DNA sequences and its application," *Chem. Phys. Lett.*, vol. 402, pp. 380–383, 2005.
[43] R. Chi and K. Ding, "Novel 4D numerical representation of DNA sequences," *Chem. Phys. Lett.*, vol. 407, pp. 63–67, 2005.
[44] B. Liao and K. Ding, "A 3D graphical representation of DNA sequences and its application," *Theor. Comput. Sci.*, pp. 56–64, 2006.
[45] X. Liu, Q. Dai, Z. Xiu, and T. Wang, "PNN-curve: A new 2D graphical representation of DNA sequences and its application," *J. Theor. Biol.*, vol. 243, pp. 555–561, 2006.
[46] Y. Zhang and W. Chen, "Invariants of DNA sequences based on 2DD-curves," *J. Theor. Biol.*, pp. 382–388, 2006.
[47] Y.-h. Yao, Q. Dai, X.-Y. Nan, P.-A. He, Z.-M. Nie, S.-P. Zhou, and Y.-Z. Zhang, "Analysis of similarity/dissimilarity of DNA sequences based on a class of 2D graphical representation," *J. Comput. Chem.*, vol. 29, pp. 1632–1639, 2008.
[48] X. Tang, P. Zhou, and W. Qiu, "On the similarity/dissimilarity of DNA sequences based on 4D graphical representation," *Chin. Sci. Bull.*, vol. 55, pp. 701–704, 2010.
[49] J. Luo, J.-C. Guo, and Y. Li, "A new graphical representation and its application in similarity dissimilarity analysis of DNA sequences," in *Proc. 4th Int. Conf. Bioinf. Biomed. Eng.*, 2010, pp. 1–5.
[50] C. Li, H. Ma, Y. Zhou, X. Wang, and X. Zheng, "Similarity analysis of DNA sequences based on the weighted pseudo-entropy," *J. Comput. Chem.*, vol. 32, pp. 675–680, 2011.

**Hong-Jie Yu** received the B.Sc. degree from Anhui University, Hefei, China, in 1996, and the M.Sc. degree from the Hefei University of Technology, Hefei, in 2004, both in mathematics. He is currently working toward the Ph.D. degree in pattern recognition and intelligent system at the University of Science and Technology of China, Hefei.

Since 1996, he has been with the Anhui Science and Technology University, Fengyang, Anhui Province, China, where he is currently an Associate Professor in the Department of Mathematics. His current research interests include intelligent computing and bioinformatics.

**De-Shuang Huang** (SM'98) received the B.Sc. degree from the Institute of Electronic Engineering, Hefei, China, the M.Sc. degree from the National Defense University of Science and Technology, Changsha, China, and the Ph.D. degree from Xidian University, Xian, China, in 1986, 1989, and 1993, respectively, all in electronic engineering.

During 1993–1997, he was a Postdoctoral Student in the Beijing Institute of Technology and in the National Key Laboratory of Pattern Recognition, Chinese Academy of Sciences, Beijing, China. In September 2000, he joined the Institute of Intelligent Machines, Chinese Academy of Sciences as the Recipient of "Hundred Talents Program of CAS." In September 2011, he joined Tongji University, Shanghai, China, as a Chaired Professor. From September 2000 to March 2001, he worked as a Research Associate at Hong Kong Polytechnic University. From August to September 2003, he visited the George Washington University as Visiting Professor, Washington, DC. From July to December 2004, he worked as the University Fellow in Hong Kong Baptist University. From March 2005 to March 2006, he worked as a Research Fellow at the Chinese University of Hong Kong. From March 2006 to July 2006, he worked as a Visiting Professor at the Queen's University of Belfast, Belfast, U.K. In 2007, 2008, 2009, he worked as a Visiting Professor at Inha University, Korea, respectively. He is currently the Head of the Machines Learning and Systems Biology Laboratory, Tongji University. He has published more than 170 journal papers. In 2001 and 2009, he published two books entitled *Intelligent Signal Processing Technique for High Resolution Radars* (in Chinese) and *The Study of Data Mining Methods for Gene Expression Profiles* (in Chinese), respectively. His current research interest includes bioinformatics, pattern recognition, and machine learning.

Dr. Huang received the Second-Class Prize of the 8th Excellent High Technology Books of China for his book *Systematic Theory of Neural Networks for Pattern Recognition* (in Chinese), published in 1996.