Tsinghua Science and Technology

Volume 18 | Issue 5 Article 3

2013

Descriptors for DNA Sequences Based on Joint Diagonalization of Their Feature Matrices from Dinucleotide Physicochemical Properties

Hongjie Yu
Department of Mathematics, School of Science, Anhui Science and Technology University, Fengyang 233100, China

Deshuang Huang
Machine Learning and Systems Biology Laboratory, Tongji University, Shanghai 201804, China

TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214 03/11 pp446-453
Volume 18, Number 5, October 2013

Descriptors for DNA Sequences Based on Joint Diagonalization of Their Feature Matrices from Dinucleotide Physicochemical Properties

Hongjie Yu and Deshuang Huang*

Abstract: Numerical characterizations of DNA sequence can facilitate analysis of similar sequences. To visualize and compare different DNA sequences in less space, a novel descriptors extraction approach was proposed for numerical characterizations and similarity analysis of sequences. Initially, a transformation method was introduced to represent each DNA sequence with dinucleotide physicochemical property matrix. Then, based on the approximate joint diagonalization theory, an eigenvalue vector was extracted from each DNA sequence, which could be considered as descriptor of the DNA sequence. Moreover, similarity analyses were performed by calculating the pair-wise distances among the obtained eigenvalue vectors. The results show that the proposed approach can capture more sequence information, and can jointly analyze the information contained in all involved multiple sequences, rather than separately, whose effectiveness was demonstrated intuitively by constructing a dendrogram for the 15 beta-globin gene sequences.

Key words: descriptors; approximate joint diagonalization; dendrogram; physicochemical property; similarity analysis

1 Introduction

Numerical characterizations of DNA sequences offer a visual means for inspection of data[1]. Thus, the description of the sequence is of great importance. However, as complex objects may be similar in one aspect and quite different in another, the space of similarity for such objects turns to be multidimensional. Recently, many numerical characterizations for DNA or protein sequences have been introduced, with most of them being extracted from string representations and graphical representations. The string representations also allowed the extraction of some simpler and more important features, which had been initially used for comparison of the genome sequence[2], and later for alignment-free comparison of regulatory sequences[3].

A survey of previous reports[4,5] revealed many frequency-based algorithms for sequence comparisons and analyses have been based on single as well as dinucleotides. Wu et al. [6] investigated the analysis approaches using adjacent nucleotides of a DNA sequence that revealed their hidden biological information whereas Randić[7] introduced a condensed representation of DNA based on pairs of nucleotides.

Some researchers[8,9] have recently introduced a graphical representation of DNA sequences based on the neighboring dinucleotides, as another example of a linear representation for strings of sequences, or have used sparse representation approach to depict objects[10]. Dual nucleotides may also be categorized into groups according to their physicochemical properties[11] (for reviews up to 2011).

In general, DNA sequences may be converted into

• Hongjie Yu is with the Department of Mathematics, School of Science, Anhui Science and Technology University, Fengyang 233100, China. E-mail: yhj70@mail.ustc.edu.cn.
• Deshuang Huang is with Machine Learning and Systems Biology Laboratory, Tongji University, Shanghai 201804, China. E-mail: dshuang@tongj.edu.cn.
∗ To whom correspondence should be addressed.
Manuscript received: 2013-06-15; revised: 2013-09-04; accepted: 2013-09-05

numerical signals through binarization processing of the sequences. Thus, the binary values 0-1 can be used to describe each two adjacent positions irrespective of their dinucleotide combination[12]. Certainly, the binary representation may be regarded as one of the earliest and the most popular transformations of DNA, although other different transformation methods[13-22] or sparse representation have also been explored.

Among these transformation methods, some are not intended for biological problems, while others have no simple numerical interpretation. Moreover, some of the representations neglect the sequence structure and the transformations tend to be irreversible. Until now, there is no ideal approach that is able to transform every type of DNA/protein sequences so as to numerically analyze the relationship among them with efficiency.

In this study, we introduce a novel method for numerical characterization of DNA sequences and apply the method to similarity analysis of these sequences, where we derived descriptors from sequences. The application of the numerical characterization of DNA sequence is illustrated by examining the relationship among different species. The results from comparison with earlier approaches demonstrated that our proposed approach is effective and is capable of analyzing the similarities among multiple sequences.

2 Descriptors of DNA Sequences

Numerical characterization of DNA sequences can facilitate us to perform similarity analysis of multiple sequences. In this section, we will introduce a novel method to transform each DNA sequence into a symmetric matrix, through which feature vectors may be readily derived.

2.1 Construction of feature matrix for sequence

Numerical characterization of 2-D or 3-D graphical representations for DNA sequences have been widely used, where some feature matrices were transformed from the primary sequences. Thus, the descriptors could be derived from these matrices. These descriptors characterizing sequence can be used as the components of similarity measures between a pair of sequences[11].

Considering a DNA sequence $S = \text{“}S_1 S_2 \cdots S_L\text{”}$, where $S_i \in \{\text{A,T,G,C}\}$, $i = 1, 2, \cdots, L$, and L denotes the length of the sequence, there are 16 kinds of dinucleotides in total (shown in Table 1). It is known that the four nucleic acids A, T, G, and C can be grouped in line with the distributions of purine-pyrimidine (R/Y), amino-keto (M/K), and weak-strong (W/S) bonds. The classifications are as follows: R = (A, G) and Y = (C, T), M = (A, C) and K = (G, T), W = (A, T) and S = (C, G).

Table 1 16 kinds of dinucleotides.

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

According to Ref. [11], many different binary techniques have been assigned the values 0-1 to Y, K, S and to R, M, W, respectively. Considering graph theory, DNA/protein sequences may be regarded as node-edge-node models, where there are four types of nodes, A-T-G-C, and 16 kinds of edges (shown in Table 1). However, all the 16 kinds of edges may be grouped into 12 kinds of dinucleotide bonds by combining every two adjacent loci.

Table 2 exhibits the decision criteria for converting dinucleotides into 12-tuple row vectors. Therefore, by scanning every adjacent site successively, such as pairs of loci (S_1, S_2), (S_2, S_3), $\cdots$, (S_{L-1}, S_L), a $(L-1)$ by 12 adjacency matrix transformed from the primary sequence may be obtained via the relationships of all the adjacent dinucleotides, referred to as $\boldsymbol{m}$:

$$\boldsymbol{m} = (\boldsymbol{\alpha}_i)_{(L-1)\times 12} \tag{1}$$

where

$$\boldsymbol{\alpha}_i = \text{the } i\text{-th row vector} \tag{2}$$

Table 2 Decision criterion for mapping dinucleotides into 0-1 values according to the combination of bonds.

$S_j S_{j+1}$	RR	RY/YR	YY	MM	MK/KM	KK	WW	WS/SW	SS
AA	1			1			1		
AT		1			1		1		
AG	1				1			1	
AC		1		1				1	
TA		1			1		1		
TT			1			1	1		
TG		1				1		1	
TC			1		1			1	
GA	1				1			1	
GT		1				1		1	
GG	1					1			1
GC		1			1				1
CA		1		1				1	
CT			1		1			1	
CG		1			1				1
CC			1	1					1

if $S_i S_{i+1}$ is the i-th kind of dinucleotide, $i = 1, 2, \cdots, L-1$.

Generally, the primary DNA/protein sequences could be considered as symbolic signals which may have a rich statistical structure. Many signal processing algorithms are focused on the organization of these statistical structures, such as the least square regression for utilizing the data correlation structure[22].

Since symmetric matrix has many merits[23], we can use the sparse matrix $\boldsymbol{m}_{(L-1)\times 12}$ mapped from the primary sequence to obtain a symmetric feature matrix $\boldsymbol{M}_{12} = \boldsymbol{m}^{\mathrm{T}} \times \boldsymbol{m}$ for representing each sequence. Furthermore, as each sequence has a different length, $\boldsymbol{M}_{12}$ could be divided by the sum of all its elements to eliminate the negative influence from non-equal length among multiple sequences. Thus, all the transformed feature matrices could become comparable.

According to signal processing, the symmetric matrix $\boldsymbol{M}_{12}$ may be interpreted as the observations from a two-layer ensemble of "sensors". The virtual sensors have an input via sixteen dinucleotides while the output is a 12×12 symmetric matrix for each DNA sequence (as depicted in Tables 1 and 2). Thus, the matrix analysis approach may be generalized from the field of signal processing to that of bioinformatics, for performing multiple sequences similarity analyses.

2.2 Feature extraction from multiple sequences via approximate joint diagonalization upon matrices

For all the multiple sequences, we can use Approximate Joint Diagonalization (AJD) of their corresponding transformed matrices $\mathcal{M} = \{\boldsymbol{M}^{(1)}, \boldsymbol{M}^{(2)}, \cdots, \boldsymbol{M}^{(N)}\}$, which has been successfully applied in the dataset with the first exon from sequences of 11 beta globin genes[24].

In brief, AJD corresponds to the problem of seeking a matrix $\boldsymbol{V}$, which will lead $\boldsymbol{V}^{\mathrm{H}}\boldsymbol{M}^{(n)}\boldsymbol{V}$ to be the diagonal as possible for all n, where $\boldsymbol{V}$ is a unitary matrix. This is based on the premise that a set of matrices $\{\boldsymbol{M}^{(1)}, \boldsymbol{M}^{(2)}, \cdots, \boldsymbol{M}^{(N)}\}$ consists of common statistical information of the observations which are the estimates of matrices in the form $\boldsymbol{V}^{\mathrm{H}}\boldsymbol{M}^{(n)}\boldsymbol{V}$.

In general, for any $n \times n$ matrix $\boldsymbol{V}$, the AJD criterion may be defined as the following non-negative function of $\boldsymbol{V}$:

$$\boldsymbol{J}(\boldsymbol{V}, \boldsymbol{\Lambda}^{(1)}, \boldsymbol{\Lambda}^{(2)}, \cdots, \boldsymbol{\Lambda}^{(N)}) \stackrel{\mathrm{def}}{=} \sum_{i=1}^{N} ||\boldsymbol{\Lambda}^{(i)} - \boldsymbol{V}^{\mathrm{H}}\boldsymbol{M}^{(i)}\boldsymbol{V}||^2 \tag{3}$$

Usually, AJD does not require that the involved matrix set $\mathcal{M}$ be exactly simultaneously diagonalized by a common unitary matrix. Mostly, the criterion for AJD, as indicated by Eq. (3), cannot be zeroed, and the matrices can only be approximately jointly diagonalized. Thus, AJD deals with a kind of an "average eigen-structure", which is particularly convenient for statistically inferring the structural information extracted from sample statistics.

Considering two transformations:

$$\tau_1 : \mathrm{Sequence}^{(i)} \mapsto \boldsymbol{M}^{(i)}.$$

$\mathrm{Sequence}^{(i)}$ denotes the i-th sequence, where the length of the sequence is L, and $i = 1, 2, \cdots, N$, while $\boldsymbol{M}^{(i)} \in \mathbb{R}^{12\times 12}$ stands for the corresponding matrices mapped from each primary DNA sequences, and $\boldsymbol{M}^{(i)}$ is a symmetric matrix, which can be determined by scanning along the sequence $S^{(i)}$ via the decision criterion listed in Table 2.

$$\tau_2 : \boldsymbol{M}^{(i)} \mapsto (\lambda_1^{(i)}, \lambda_2^{(i)}, \cdots, \lambda_{12}^{(i)}).$$

The feature vector $\boldsymbol{F}_{12}^{(i)} = (\lambda_1^{(i)}, \lambda_2^{(i)}, \cdots, \lambda_{12}^{(i)})$ is a 12-tuple vector consisting of all the eigenvalues extracted by AJD upon $\boldsymbol{M}^{(i)}$. Thus, compound transformation may be obtained as follows:

$$\tau_2 \circ \tau_1 : \mathrm{Sequence}^{(i)} \mapsto (\lambda_1^{(i)}, \lambda_2^{(i)}, \cdots, \lambda_{12}^{(i)}) \tag{4}$$

From Formula (4), we can freely extract the features of the DNA sequence. From the viewpoint of algebra space, the transformation may also be presented as

$$\mathrm{Ker}\ f : \boldsymbol{S} \xrightarrow{\tau} \boldsymbol{F}^{1\times 12} \tag{5}$$

where $\boldsymbol{S}$ denotes the original sequence space comprising of primary DNA sequence having the length L, while $\boldsymbol{F}^{1\times 12}$ indicates the *objective feature space* that is transformed from the original space. Further, the diagonal elements of $\boldsymbol{\Lambda}$ are simply the eigenvalues of the dinucleotide PhysicoChemical Matrix (PCM) via AJD.

According to the results from previous work[24], it was determined that the AJD-PCM algorithm has the property of distance-preserving. Thus, we can calculate all the Eigenvalue Vectors (EVs) for each obtained dinucleotide PCM, such as $\boldsymbol{F}_{12}^{(i)} = (\lambda_1^{(i)}, \lambda_2^{(i)}, \cdots, \lambda_{12}^{(i)})$, $i = 1, 2, \cdots, N$. Further, the N corresponding 12-tuple vectors may be obtained that can be regarded as features extracted from the original DNA sequence. The AJD-PCM algorithm is shown in Algorithm 1.

3 Numerical Characterization of Sequences

Most numerical characterization methods can represent each sequence only *separately* rather than *jointly*. In this

Algorithm 1 Descriptors for DNA sequences via AJD-PCM

Input: Multiple DNA sequences with different length $L, S^{(1)}, S^{(2)}, \cdots, S^{(N)}$.

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 $S^{(n)}$ into 12 by 12 symmetric matrix $\boldsymbol{M}^{(n)}$
  end for
- Consider the obtained matrix set $\mathcal{M} = \{\boldsymbol{M}^{(1)}, \boldsymbol{M}^{(2)}, \cdots, \boldsymbol{M}^{(N)}\}$ and objective function
  $$\boldsymbol{J}(\boldsymbol{V}, \boldsymbol{\Lambda}^{(1)}, \boldsymbol{\Lambda}^{(2)}, \cdots, \boldsymbol{\Lambda}^{(N)}) \stackrel{\text{def}}{=} \sum_{i=1}^{N} ||\boldsymbol{\Lambda}^{(i)} - \boldsymbol{V}^{\text{H}} \boldsymbol{M}^{(i)} \boldsymbol{V}||^2.$$
  while $\boldsymbol{J}(\boldsymbol{V}, \boldsymbol{\Lambda}^{(1)}, \boldsymbol{\Lambda}^{(2)}, \cdots, \boldsymbol{\Lambda}^{(N)}) \geqslant$ tolerance **do**
  Update $\boldsymbol{V}$ using AC-DC algorithm[25]
  end while
- **for** $n = 1$ **to** N **do**
  $\boldsymbol{F}^{(n)} \leftarrow \text{diag}(\boldsymbol{\Lambda}^{(n)})$
  Plot and categorize the N feature curves with $\boldsymbol{F}^{(n)}$
  end for
- Calculate pairwise distances using all these N vectors $\boldsymbol{F}^{(n)}$
- Draw the dendrogram using the pairwise distances matrix

end

section, a novel method of numerical characterization and graphical representation for DNA sequences is presented, which can *jointly* consider the mutual information of all the involved sequences. We selected a dataset comprising of 15 sequences of the first exon in the *beta*-globin gene[26], as shown in Table 3.

3.1 Calculation of eigenvalue vectors

To enable comparison of multi-sequences, all matrices extracted from each corresponding sequence were normalized by dividing each $\boldsymbol{M}_{12}$ with the sum of values for all their elements. According to the procedure for AJD-PCM algorithm depicted in Section 2.2, all the 12-tuple EVs are calculated through AJD of all the eleven neighboring nucleotide matrices (PCM). The results are listed in Table 4, where each column denotes a 12-tuple EV for each sequence. If each vector is connected in an orderly manner as head and tail using the walk strategy, 15 curves may be plotted, as shown in Fig. 1.

Table 3 The concise information from beta-globin genes of 15 species.

Species	AC (GeneBank)	Location	Length (nt)
Human	U01317	62 187-63 610	1424
Chimpanzee	X02345	4189-5532	1344
Gorilla	X61109	4538-5881	1344
Lemur	M15734	154-1595	1442
Rat	X06701	310-1505	1196
Mouse	V00722	275-1462	1188
Goat	M15387	279-1749	1471
Sheep	DQ352470	238-1708	1471
Mouflon	DQ352468	238-1706	1469
Bovine	X00376	278-1741	1464
Rabbit	V00882	277-1419	1143
European hare	Y00347	1485-2620	1136
Opossum	J03643	467-2488	2022
Gallus	V00409	465-1810	1346
Muscovy duck	X15739	291-1870	1580

3.2 Convergence speed analysis

Based on the convergence analysis of AJD, the convergence speed of the optimization scheme (see Eq. (3)) should be considered. The diagonalizer matrix $\boldsymbol{V}$ of the objective function at the optimum point depends upon an unknown least error. Thus, a simplified expression for the least error is as follows:

$$\text{Err}(j) \stackrel{\text{def}}{=} \sum_{i=1}^{N} ||\boldsymbol{\Lambda}_j^{(i)} - \boldsymbol{V}_j^{\text{H}} \boldsymbol{M}^{(i)} \boldsymbol{V}_j||_{\text{F}}^2 \qquad (6)$$

where j ranges from 2 to a presupposed maximal number of iterations.

The predefined error threshold is denoted as ϵ when $|\text{Err}(j+1) - \text{Err}(j)| < \epsilon$, which indicates that when the error does not change drastically from the j-th iteration to the $(j+1)$-th one, we can get either the obtained diagonalizer $\boldsymbol{V}$ or the diagonalized $\boldsymbol{\Lambda}^{(i)}$, which just occurs at j-th step. Figure 2 shows a plot of the errors with each iteration, from which it could be found that the AJD algorithm converged at just the third iteration or so.

4 Similarity Analysis of Sequences Based on Their Extracted Descriptors

Based on sequence descriptors extracted via AJD-PCM, multiple sequences can be compared.

4.1 Calculation of pair-wise distances

Consequently, numerical characterization was applied for investigating the similarity of multiple sequences from the dataset listed in Table 3. As described in Section 2.2, the genetic distance was calculated in the present study. For every two EVs: $\boldsymbol{F}^{(i)}$ and $\boldsymbol{F}$, which have been arranged within the corresponding columns in Table 4, it was found that the degree of dissimilarity may be determined through Euclidean distance between every two column vectors. The Euclidean distance between the i-th and the j-th sequences may be

Table 4 Descriptors extracted via AJD-PCM from the beta-globin genes of 15 species.

No.	Human	Chimpanzee	Gorilla	Lemur	Rat	Mouse	Goat	Sheep	Mouflon	Bovine	Rabbit	European hare	Opossum	Gallus	Muscovy duck
1	0.019 91	0.019 41	0.019 65	0.021 09	0.022 71	0.021 86	0.025 50	0.025 51	0.025 67	0.025 63	0.021 34	0.021 68	0.023 12	0.030 92	0.032 09
2	0.024 29	0.024 15	0.024 13	0.024 07	0.023 47	0.023 35	0.023 55	0.023 55	0.023 50	0.022 85	0.024 18	0.024 80	0.020 42	0.024 19	0.023 40
3	0.027 06	0.027 33	0.027 43	0.026 30	0.025 23	0.025 16	0.024 35	0.024 28	0.024 31	0.024 03	0.025 31	0.025 89	0.023 78	0.019 84	0.018 44
4	0.030 13	0.031 15	0.030 93	0.030 71	0.031 76	0.032 74	0.028 85	0.028 73	0.028 39	0.028 62	0.032 13	0.030 88	0.034 51	0.019 62	0.019 49
5	0.023 77	0.022 11	0.022 11	0.022 48	0.026 44	0.025 98	0.025 41	0.025 74	0.025 87	0.026 29	0.024 16	0.023 81	0.026 36	0.032 74	0.033 50
6	0.025 81	0.026 12	0.026 33	0.025 53	0.028 51	0.028 84	0.029 78	0.029 32	0.029 24	0.028 69	0.026 44	0.027 31	0.029 49	0.026 48	0.026 70
7	0.021 47	0.021 22	0.021 68	0.022 01	0.024 22	0.025 65	0.027 06	0.026 71	0.026 85	0.026 72	0.023 46	0.024 08	0.024 48	0.028 40	0.029 19
8	0.036 47	0.037 88	0.036 53	0.043 67	0.034 89	0.033 09	0.026 95	0.027 46	0.026 61	0.027 02	0.041 14	0.035 94	0.022 18	0.030 33	0.024 30
9	0.041 68	0.044 36	0.044 52	0.034 47	0.025 42	0.027 07	0.018 95	0.019 40	0.019 52	0.020 77	0.029 37	0.029 21	0.035 77	**−0.012 66**	**−0.013 26**
10	0.028 49	0.028 50	0.028 47	0.029 54	0.031 32	0.030 46	0.031 59	0.031 72	0.031 57	0.030 96	0.030 85	0.031 16	0.030 30	0.03171	0.031 88
11	0.028 96	0.028 64	0.028 87	0.029 41	0.030 42	0.029 78	0.030 97	0.031 10	0.031 10	0.030 78	0.031 04	0.031 75	0.029 74	0.03334	0.033 19
12	0.018 45	0.017 79	0.017 32	0.020 02	0.019 93	0.021 32	0.021 91	0.021 11	0.021 14	0.021 74	0.019 55	0.019 37	0.017 30	0.038 27	0.037 51

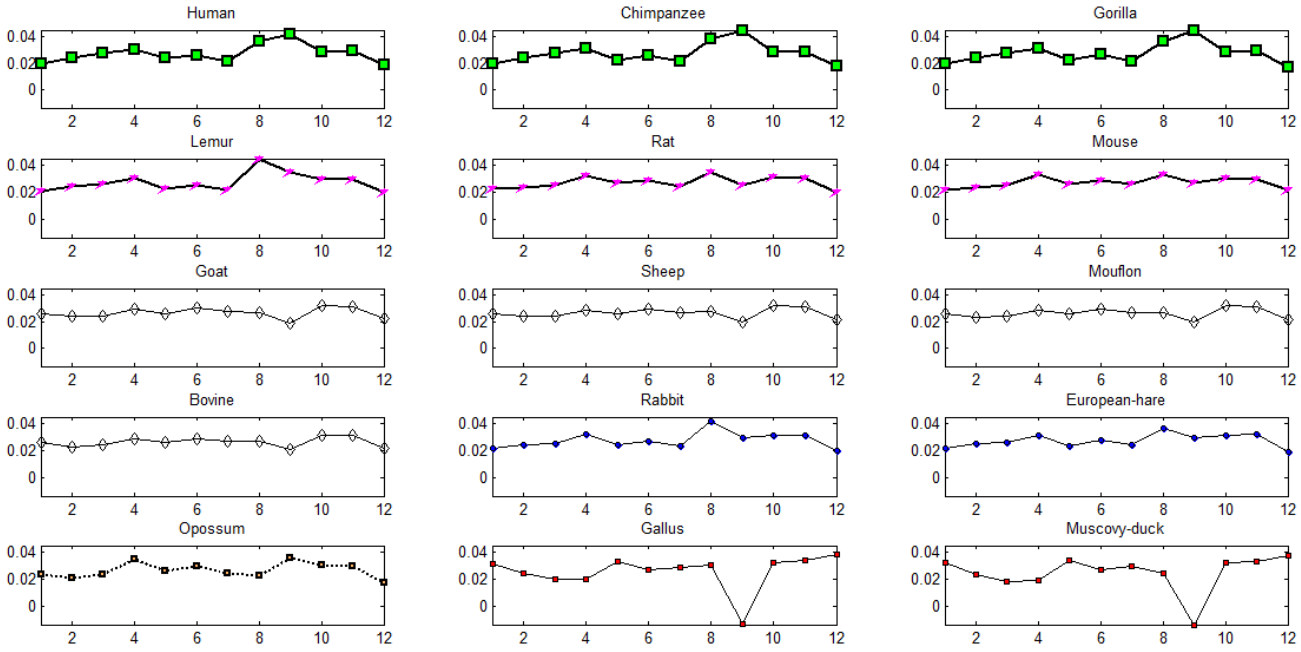

Fig. 1 Graphical representation of the exon in the beta-globin gene from 15 species based on 12-tuple EVs via AJD upon all the corresponding PCP matrices. The *y*-axis indicates the values of each element in feature vectors $(\lambda_1, \lambda_2, \cdots, \lambda_{12})$.

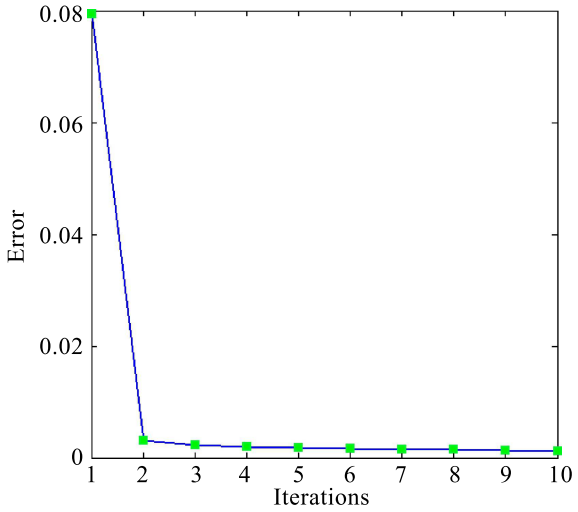

Fig. 2 The performance curve for the convergence speed of AJD for the dataset listed in Table 3.

calculated as:

$$D(S^{(i)}, S^{(j)}) \stackrel{\text{def}}{=} ||\boldsymbol{M}^{(i)} - \boldsymbol{M}^{(j)}||_{\text{F}} = ||\boldsymbol{F}^{(i)} - \boldsymbol{F}^{(j)}||_{\text{F}} \tag{7}$$

where $\boldsymbol{F}^{(i)} = (\lambda_1^{(i)}, \lambda_2^{(i)}, \cdots, \lambda_{12}^{(i)})$ denotes the feature vectors through distance-preserving transformations from the primary sequences with different lengths, while $||\cdot||_{\text{F}}$ indicates the Frobenius norm of a matrix or vector. Obviously, a higher distance value indicates more dissimilarity between the two sequences. The data for comparison of dissimilarity among these 15 coding sequences was obtained by calculating the Euclidean distance via Eq. (7). The pair-wise distance results are listed in Table 5.

4.2 Phylogeny of 15 beta-globin genes

Other than the alphabet representation of biological sequences, which is easily processed on a computer, it

Table 5 Pairwise distance for the beta-globin genes of 15 species.

Species $i \setminus j$	Chimpanzee	Gorilla	Lemur	Rat	Mouse	Goat	Sheep	Mouflon	Bovine	Rabbit	European hare	Opossum	Gallus	Muscovy duck
Human	0.0037	0.0036	0.0106	0.0178	0.0169	0.0269	0.0262	0.0264	0.0252	0.0141	0.0136	0.0181	0.0619	0.0636
Chimpanzee		0.0016	0.0120	0.0208	0.0199	0.0301	0.0294	0.0297	0.0285	0.0164	0.0165	0.0204	0.0652	0.0670
Gorilla			0.0128	0.0207	0.0197	0.0297	0.0290	0.0292	0.0280	0.0168	0.0165	0.0193	0.0652	0.0669
Lemur				0.0141	0.0146	0.0247	0.0240	0.0245	0.0235	0.0068	0.0103	0.0234	0.0562	0.0588
Rat					0.0036	0.0117	0.0109	0.0116	0.0107	0.0082	0.0055	0.0172	0.0461	0.0478
Mouse						0.0118	0.0112	0.0117	0.0106	0.0095	0.0060	0.0151	0.0472	0.0488
Goat							0.0013	0.0014	0.0025	0.0193	0.0153	0.0196	0.0385	0.0392
Sheep								0.0010	0.0021	0.0185	0.0145	0.0192	0.0391	0.0398
Mouflon									0.0019	0.0192	0.0151	0.0190	0.0391	0.0397
Bovine										0.0182	0.0142	0.0180	0.0398	0.0405
Rabbit											0.0056	0.0212	0.0512	0.0536
European hare												0.0171	0.0499	0.0518
Opossum													0.0570	0.0571
Gallus														0.0065

becomes difficult to observe other differences without aid, for which a phylogenetic tree provides a simpler means to inspect various biological sequences that facilitates sequence comparison with the intuitive patterns or pictures. Thus, we examined our proposed approach (AJD-PCM) through phylogenetic analysis, which may be summarized as follows:

(1) Initially, the 12-tuple EVs of each biological sequence were calculated through AJD-PCM;

(2) Then, the similarity distance was obtained with the Euclidean metric to form a pair-wise distance matrix;

(3) Finally, based on the pair-wise distance matrix, the dendrogram was plotted using MATLAB code.

The results of the experiment have been listed in Table 5 and Fig. 3, from which it was found that the 15 species were clearly separated.

Figure 3 displays the intuitive categories of the 15 species into four subgroups according to the divergence between themselves and humans as follows:

(1) The first subgroup [(Gorilla, Chimpanzee), Human] is closest to the humans.

(2) The next subgroup includes the species [Bovine, (Goat, (Mouflon, Sheep))].

(3) The third ((Muscovy duck, Gallus), Opossum) is far away from the humans in the light of evolutionary relationships. Generally, these three species may be regarded as outgroup.

(4) The remaining five species fall into the last group.

These results are consistent with the evolutionary facts.

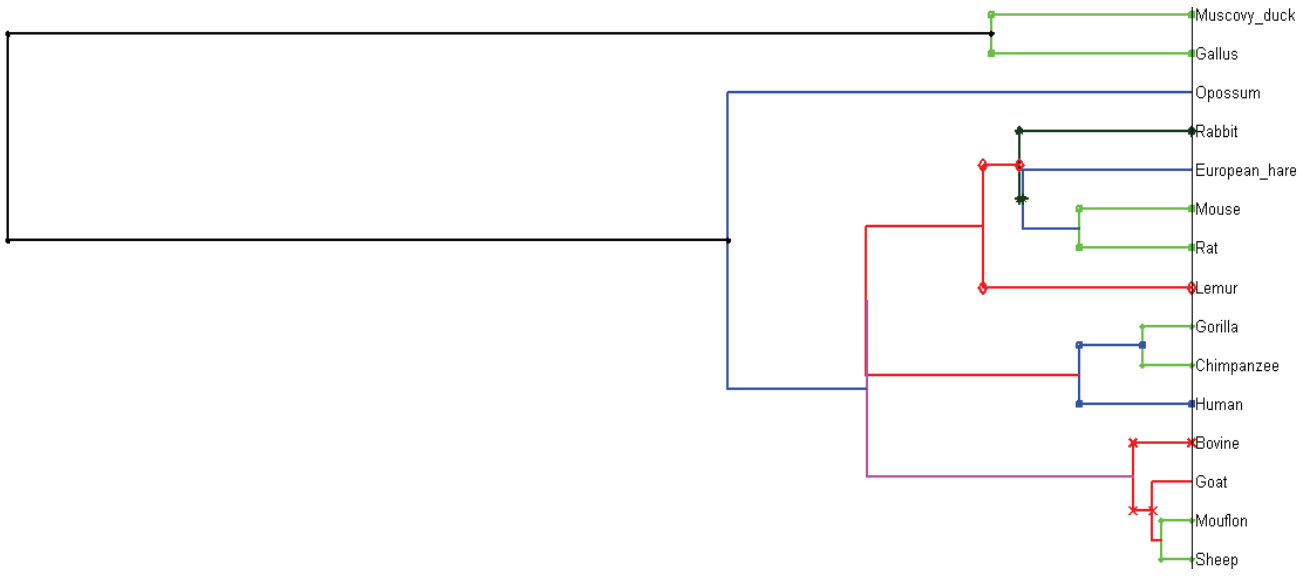

Fig. 3 The dendrogram for 15 sequences according to the pairwise distance listed in Table 3.

4.3 Comparison with related representative work

To comprehend the effectiveness of our approach, we compared the pair-wise distances obtained from the present study (Table 5) with those in the seventh table from another related representative work[27]. Considering the similarity among Humans and the remaining 14 species, we extracted the first row (or column) from the two tables mentioned above, and calculated the correlation coefficient for these two reduced vectors, that is, the first row of Table 5 in this study, and the first column from the seventh table from Ref. [27]. The correlation degree was determined to be 98.76%. Remarkably, comparison of the two whole tables above-mentioned shows the correlation degree also reaches up to 93.39%.

5 Conclusions and Future Work

The proposed approach (AJD-PCM) allows the fusion of the physicochemical properties with the sequential property of the biological sequence with consideration of the sequential property at the stage of mapping each dinucleotide into a 12-tuple vector. Moreover, at the second stage, AJD could extract the features among multiple sequences *jointly* rather than *separately*, facilitating the simultaneous discovery of sub-groups of organisms having a common structure at the molecular level. The clustering results were consistent with the evolutionary facts demonstrating the rationality of the proposed numerical characterization of DNA sequence. Further work would involve enhancing our descriptor extraction algorithms for application towards genomic/proteomic sequence datasets.

Acknowledgements

This work was supported by the Key Project from Education Department of Anhui Province (No. KJ2013A076), the PhD Programs Foundation of Ministry of Education of China (No. 20120072110040), the National Natural Science Foundation of China (Nos. 61133010, 31071168, and 61005010), and the China Postdoctoral Science Foundation (No. 2012T50582).

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] B. E. Blaisdell, A measure of the similarity of sets of sequences not requiring sequence alignment, *Proceedings of the National Academy of Sciences*, vol. 83, pp. 5155-5159, 1986.

[3] M. R. Kantorovitz, G. E. Robinson, and S. Sinha, A statistical method for alignment-free comparison of regulatory sequences, *Bioinformatics*, vol. 23, no. 13, pp. i249-i255, 2007.

[4] 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, *Proceedings of the National Academy of Sciences*, vol. 106, no. 8, pp. 2677-2682, 2009.

[5] 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, *Proceedings of the National Academy of Sciences*, vol. 107, no. 1, pp. 133-138, 2009.

[6] Y. Wu, A. W-C Liew, H. Yan, and M. S. Yang, DB-Curve: A novel 2D method of DNA sequence visualization and representation, *Chemical Physics Letters*, vol. 367, pp. 170-176, 2003.

[7] M. Randić, Condensed representation of DNA primary sequences, *Journal of Chemistry Information & Computer Science*, vol. 40, no. 1, pp. 50-56, 2000.

[8] 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.

[9] Z. B. Liu, B. Liao, W. Zhu, and G. H. Huang, A 2D graphical representation of DNA sequence based on dual nucleotides and its application, *International Journal of Quantum Chemistry*, vol. 109, no. 5, pp. 948-958, 2009.

[10] C. Y. Lu, H. Min, J. Gui, L. Zhu, and Y. K. Lei, Face recognition via weighted sparse representation, *Journal of Visual Communication and Image Representation*, vol. 24, no. 2, pp. 111-116, 2013.

[11] D. Bielinska-Waz, Graphical and numerical representations of DNA sequences: Statistical aspects of similarity, *Journal of Mathematical Chemistry*, vol. 49, no. 10, pp. 2345-2407, 2011.

[12] R. F. Voss, Evolution of long-rang fractal correlations and 1/f noise in DNAbase sequences, *Physical Review Letter*, vol. 68, pp. 3805-3808, 1992.

[13] M. Akhtar, J. Epps, and E. Ambikairajah, On DNA numerical representation for period-3 based exon prediction, in 5^{th} *International Workshop on Genomic Signal Processing and Statistics, Tuusula*, Piscataway, NJ, USA, 2007.

[14] H. J. Jeffrey, Chaos game representation of gene structure, *Nucleic Acids Research*, vol. 18, no. 8, pp. 2163-2170, 1990.

[15] C. Y. Lu and D. S. Huang, Optimized projections for sparse representation based classification, *Neurocomputing*, vol. 113, pp. 213-219, 2013.

[16] R. Zhang and C. T. Zhang, Z curves, an intutive tool for visualizing and analyzing the DNA sequences, *Journal of Biomolecular Structure & Dynamics*, vol. 11, no. 4, pp. 767-782, 1994.

[17] M. Randić, Another look at the chaos-game representation of DNA, *Chemical Physics Letters*, vol. 456, no. 1, pp. 84-88, 2008.

[18] S. Wang, F. Tian, W. Feng, and X. Liu, Applications of representation method for DNA sequences based on symbolic dynamics, *Journal of Molecular Structure: THEOCHEM*, vol. 909, pp. 33-42, 2009.

[19] A. K. Brodzik and O. Peters, Symbol-balanced quaternionic periodicity transform for latent pattern detection in DNA sequences, in *Proceedings of IEEE ICASSP*, Philadelphia, PA, USA, 2005, pp. 373-376.

[20] B. Liao, M. Tan, and K. Ding, Application of 2-D graphical representation of DNA sequence, *Chemical Physics Letters*, vol. 414, pp. 296-300, 2005.

[21] W. Wang and D. H. Johnson, Computing linear transforms of symbolic signals, *IEEE Transactions on Signal Processing*, vol. 50, no. 3, pp. 628-634, 2002.

[22] C. Y. Lu, H. Min, Z. Z. Zhao, L. Zhu, D. S. Huang, and S. C. Yan, Robust and efficient subspace segmentation via least squares regression, *European Conference on Computer Vision ECCV*, vol. 7578, no. 7, pp. 347-360, 2012.

[23] G. H. Golub and C. F. V. Loan, *Matrix Computations*, 3rd Ed. Baltimore and London: Johns Hopkins University Press, 1996.

[24] H. J. Yu and D. S. Huang, Graphical representation for DNA sequences via joint diagonalization of matrix pencil, *IEEE Journal of Biomedical and Health Informatics*, vol. 17, no. 3, pp. 503-511, 2013.

[25] A. Yeredor, Non-orthogonal joint diagonalization in the least-squares sense with application in blind source separation, *IEEE Transactions on Signal Processing*, vol. 50, no. 7, pp. 1545-1553, 2002.

[26] Q. Dai, X. Q. Liu, Y. H. Yao, and F. K. Zhao, Sequence comparison via polar coordinates representation and curve tree, *Journal of Theoretical Biology*, vol. 292, pp. 78-85, 2011.

[27] C. Li, H. Ma, Y. Zhou, X. L. Wang, and X. Q. Zheng, Similarity analysis of DNA sequences based on the weighted pseudo-entropy, *Journal of Computational Chemistry*, vol. 32, no. 4, pp. 675-680, 2011.

Hongjie Yu received his BS degree in mathematics form Anhui University, China in 1996, his MS degree in mathematics from Hefei University of Technology, China in 2004, and his PhD degree in pattern recognition and intelligent system from University of Science and Technology of China in 2013. Since 1996, he has been with the Anhui Science and Technology University, Fengyang, Anhui Province, China, where he is now an associate professor in the Department of Mathematics. His current research interests include bioinformatics, biostatistics, machine learning, and intelligent computing.

Deshuang Huang received his MS and PhD degrees in electronic engineering from the National Defense University of Science and Technology and Xidian University, China, in 1989 and 1993, respectively. He is a chaired professor in Department of Computer Science and the director of Machine Learning and Systems Biology Laboratory at Tongji University, China. His primary research interests include neural networks, pattern recognition, and bioinformatics. He is currently a Bioinformatics and Bioengineering/Neural Networks Technical Committee member of IEEE CIS, a member of the International Neural Network Society (INNS), a board member of the INNS Governors, and an associated editor of several mainstream international journals such as Neural Networks. He has published more than 300 papers in international journals, international conferences proceedings, and book chapters. He also published three monographs (in Chinese), one of which, entitled Systematic Theory of Neural Networks for Pattern Recognition, won the Second-Class Prize of the Eighth Excellent High Technology Books of China in 1997. He is currently a senior member of the IEEE.