Journal of Theoretical Biology 363 (2014) 145–150

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

k-mer Sparse matrix model for genetic sequence and its applications in sequence comparison

Jia Wen [a,*], YuYan Zhang [b], Stephen S.T. Yau [c,*]

[a] School of Information Engineering, Suihua University, Suihua 152061, PR China
[b] School of Agriculture and Hydraulic Engineering, Suihua University, Suihua 152061, PR China
[c] Department of Mathematical Sciences, Tsinghua University, Beijing 100084, PR China

HIGHLIGHTS

- *k*-mer Sparse matrix denotes the types and sites of *k*-mers in genetic sequence.
- Relationship of genetic sequence and its associated *k*-mer sparse matrix is one-to-one.
- *k*-mer Singular value vector is numerically characterized genetic sequence.
- Accuracy of sequence comparison is improved by our proposed method.

ARTICLE INFO

Article history:
Received 6 April 2014
Received in revised form
14 July 2014
Accepted 17 August 2014
Available online 23 August 2014

Keywords:
k-mer Model
Singular value decomposition
Optimum value
Phylogenetic analysis

ABSTRACT

Based on the *k*-mer model for genetic sequence, a *k*-mer sparse matrix representation is proposed to denote the types and sites of *k*-mers appearing in a genetic sequence, and there exists a one-to-one relationship between a genetic sequence and its associated *k*-mer sparse matrix. With the singular value decomposition of the *k*-mer sparse matrix, the *k*-mer singular value vector is constructed and utilized to numerically quantify the characteristics of a genetic sequence. We investigate and evaluate the optimum value k^* chosen for our *k*-mer sparse matrix model for genetic sequence. To show the usefulness of our *k*-mer sparse matrix model method, it is applied to the comparison of genetic sequences, and the results obtained fully demonstrate that our proposed method is very powerful in analyzing and determining the relationships of genetic sequences.

1. Introduction

Sequence comparison of genetic sequences is a basic but very important task in analyzing and determining the similarities/dissimilarities of genetic sequences. Based on the similarity analysis in sequence comparison, we can infer and predict the structure and function of biological molecules, and explore and clarify the evolutionary relationships of all types of individual organisms, etc. Until now, many powerful methods have been proposed for the comparison of genetic sequences. Conventional methods for sequence comparison are alignment-based methods, which are evaluated within well-established framework of alignment. Although most alignment-based methods could precisely reflect the relationships of genetic sequences, they are frequently accompanied with very high computational complexity, especially for multiple sequence alignment. Additionally, alignment-based methods have difficulty in dealing with large database and choosing scoring schemes for different kinds of genetic sequences considered. Hence, the alignment-free approach based on the numerical characterizations of genetic sequence is desirable to compensate for the ineffectiveness of alignment-based methods. Among all alignment-free methods, the *k*-mer model method may be the best developed one. Because the *k*-mer mode method is a frequency-based method, it is much faster and therefore widely used in the comparison of whole genomes. A classic *k*-mer model method was proposed for the comparison of genome sequences by Blaisdell (1986), and the counts of *k*-mers appearing in the sequence were used for the comparison of regulatory sequences by Kantorovitz et al. (2007). Various *k*-mer model methods were proposed for sequence comparison by Wu et al. (1997, 2001, 2005), Korf and Rose (2009), Sims et al. (2009a, 2009b), Jun et al. (2010), Dai et al. (2011), Yu (2013) and Yang and Wang (2013).

One of the most important but also very difficult problems in computational biology is how to formulate a biological sequence by a vector or matrix, yet still keeping considerable sequence-order information. To avoid completely losing the sequence-order

* Corresponding authors. Tel.: +86 10 62787874; fax: +86 10 62798033.
E-mail addresses: wenjia198021@126.com (J. Wen), yau@uic.edu (S.S.T. Yau).

http://dx.doi.org/10.1016/j.jtbi.2014.08.028
0022-5193/

information for protein/peptide sequence, the pseudo amino acid composition (PseAAC) was proposed by Chou (2001, 2005). In PseAAC, the sequence-order information of a protein/peptide sequence is approximately reflected by a series of correlation factors that form the components of a vector. Ever since the concept of PseAAC was proposed in 2001, it has been widely used in analyzing varieties of problems for protein/peptide sequence (Xu et al., 2013; Nanni et al., 2012; Mondal and Pai, 2014; Zou et al., 2011; Hayat and Khan, 2012; Khosravian et al., 2013; Mohabatkar et al., 2013; Mohammad et al., 2011; Nanni and Lumini, 2008; Esmaeili et al., 2010; Mohabatkar, 2010; Hajisharifi et al., 2014). Stimulated by the successes of using PseAAC to deal with protein/peptide sequence, recently the pseudo k-tuple nucleotide composition (PseKNC) was developed to represent DNA/RNA sequence for studying some important problems in genome/genetic analyses (Chen et al., 2012, 2013, 2014; Qiu et al., 2014; Guo et al., 2014; Yang et al., 2013; Gao and Luo, 2012; Cheng et al., 2013). Based on the k-mer model of genetic sequence, we are to propose a new k-mer sparse matrix model method in the hope that it will become a useful tool for analyzing genetic sequence.

2. Materials and methods

According to a recent comprehensive review by Chou (2011) and also demonstrated by a series of recent publications (Lin et al., 2011; Chou and Shen, 2010; Xiao et al., 2011; Chou et al., 2012; Chen et al., 2012, 2013; Guo et al., 2014; Feng et al., 2013), to establish a really useful statistical method for a biological system, we need to consider the following procedures: (i) construct or select a valid benchmark dataset to train and test the model; (ii) formulate the biological sample with an effective mathematical expression that can truly reflect its intrinsic correlation with the target to be analyzed; (iii) introduce or develop a powerful algorithm (or engine) to operate the analysis; (iv) properly perform testing methods to objectively evaluate the anticipated accuracy of the statistical method; (v) establish a user-friendly web-server for the method that is accessible to the public. Below, let us describe how to deal with these steps.

2.1. k-mer Sparse matrix representation

The sparse matrix representation for protein primary sequence was proposed by Yu and Huang (2012). Similarly, we apply the sparse matrix representation to the k-mer model of genetic sequence.

Let s be a genetic sequence of length L, $s=N_1N_2\ldots N_L$, where $N_l\in\{A,C,G,T\}, l=1,2,\ldots,L$. Any consecutive $k(k\geq 1)$ letters within a sequence is called a k-mer, and there would be 4^k different possible k-mers appearing in a genetic sequence. For any given k, there are $L-k+1$ k-mers in the sequence s.

The k-mer sparse matrix M for the sequence s is a 4^k by $L-k+1$ matrix, which can be obtained by using a sliding window of length k, shifting the frame one letter each time from position 1 to $L-k+1$, until the entire sequence has been scanned. For step $j, (j=1,2,\ldots, L-k+1)$, if k letters in the frame are just the ith $(i=1,2,3,\ldots,4^k)$ k-mer among 4^k different k-mers, the element in M is marked $M_{ij}^{(s)}=1$, and other elements in column j are zeroes.

It is easy to find that there would be only one 1 in each column of the k-mer sparse matrix M, and the sum of each row equals the number of times the corresponding k-mer appearing in the sequence s. Thus, for any given k, a genetic sequence can be represented by a k-mer sparse matrix, which denotes the types and sites of k-mers appearing in the sequence, and there exists a one-to-one relationship between a genetic sequence and its associated k-mer sparse matrix. The k-mer sparse matrix contains all the information hidden in a genetic sequence.

2.2. k-mer Singular value vector

To numerically characterize a genetic sequence, some key features should be obtained from the corresponding k-mer sparse matrix. Many powerful tools can be utilized to extract characteristic information from the k-mer sparse matrix. The singular value decomposition of the k-mer sparse matrix is one of the available tools, and the singular value vector composed of singular values can be used to numerically depict the characteristics of a genetic sequence.

The singular value decomposition (SVD) of the k-mer sparse matrix M can be written as follows:

$$M=U\Sigma V^T, \tag{1}$$

where U is a real square matrix satisfying $U^TU=I_{4^k}$, (I_{4^k} is a unit matrix of rank 4^k), Σ is a $4^k\times(L-K+1)$ diagonal matrix with nonnegative singular values $\sigma_1,\sigma_2,\ldots,\sigma_{k'=\min(4^k,L-K+1)}$ on the diagonal, and V^T (the transpose of V) is a real square matrix having $V^TV=I_{L-k+1}$.

By the above definition, we note that the number of singular values is relevant to the value of the given k and the length L of the sequence considered. Because the lengths of genetic sequences are commonly different, the dimension of the singular value vector is not unique especially when 4^k is larger than $L-k+1$. To facilitate the comparison of genetic sequences by some distance measure, it is better to obtain the same dimensional singular value vector for sequences of different lengths when k is fixed. Hence, we do a little changing of the k-mer sparse matrix when 4^k is larger than $L-k+1$: the k-mer sparse matrix would be extended to a square matrix of rank 4^k by means of a $4^k\times(4^k-(L-k+1))$ zero matrix added to the end of k-mer sparse matrix. By this change, we always obtain 4^k singular values by the SVD of k-mer sparse matrices for genetic sequences of different lengths when k is given. It is nice that not only all information hidden in the genetic sequences is well reserved, but also the same dimensional singular value vectors can be gotten to characterize genetic sequences of different lengths.

Therefore, for each fixed k, the k-mer singular value vector can be constructed and utilized to numerically depict the characteristics of a genetic sequence, in which 4^k singular values are sorted in the lexicographic order of k-mers among the singular value vector.

2.3. The optimum value k* for k-mer sparse matrix model

Since the parameter k has a great influence on computational complexity and the result of sequence comparison, it is vital to choose an appropriate k for our k-mer sparse matrix model. It has been shown that it is important and difficult to choose a suitable k for genetic sequences of different lengths considered in the k-mer model. Some researchers have investigated the selection of the optimum value k^* for the k-mer model. For example, Wu et al. (2005) pointed out that an optimal word size k^* for dissimilarity measurement should be increased when the sequence length increases, and Sims et al. (2009a, 2009b) reported that the optimal word length lies within an approximate range with lower bound $\log_4 n$, where n is the length of sequence, and upper bound given by the criterion that the phylogenetic tree topology for length k must be parallel to that of $k+1$.

Until now, there is no recognized criterion on choosing the optimum k^* for k-mer model. The optimum value k^* for our k-mer sparse matrix model is chosen when the result of sequence comparison obtained for a given k is consistent with the standard

biological taxonomy, the evolutionary relationships of species, and results obtained by sequence alignment and some published papers. Following former work in searching the optimum value for k-mer model, we infer that the optimum value k^* for our k-mer sparse matrix model is close to the value of floor $(\log_4 \text{mean}(L))$, where L is the length set of genetic sequences considered.

Although the larger k is the main obstacle for our k-mer sparse matrix model method, our proposed method is faster than multiple sequence alignment methods. The runtime comparisons for our k-mer sparse matrix method, ClustalW and MUSCLE are shown in Appendix A. One needs to do multiple sequence alignment again entirely when one adds one more sequence in the dataset considered. However, for our proposed method all the singular value vectors can be stored and we do not need to recompute them.

2.4. Distance metric

Since each genetic sequence can be uniquely represented by its k-mer singular value vector, a distance metric can be used to quantify the similarities/dissimilarities of genetic sequences. The similarity between each pair of genetic sequences can be computed by the correlation angle of their k-mer singular value vectors, because the correlation angle can eliminate the effects of high dimensionality (Berry et al., 1999; Wen and Zhang, 2009; Wen et al., 2014). In this paper, we select the distance metric defined below to measure the similarities/dissimilarities of genetic sequences. This distance matric has been widely used in k-mer models (Qi et al., 2004; Stuart et al., 2002; Stuart and Berry, 2004).

Let v_1 and v_2 be the k-mer singular value vectors of two genetic sequences s_1 and s_2, respectively, the distance between sequences s_1 and s_2 can be computed as follows:

$$d(s_1, s_2) = 1 - \cos(v_1, v_2) = 1 - \frac{v_1 \times v_2}{|v_1||v_2|}, \quad (2)$$

Table 1
Description of 31 mitochondrial genome sequences.

No.	Individual	GenBank ID
1	Human	V00662
2	Pigmy chimpanzee	D38116
3	Common chimpanzee	D38113
4	Gibbon	X99256
5	Baboon	Y18001
6	Vervet monkey	AY863426
7	Macaca thibetana	NC_002764
8	Bornean orangutan	D38115
9	Sumatran orangutan	NC_002083
10	Gorilla	D38114
11	Cat	U20753
12	Dog	U96639
13	Pig	AJ002189
14	Sheep	AF010406
15	Goat	AF533441
16	Cow	V00654
17	Buffalo	AY488491
18	Wolf	EU442884
19	Tiger	EF551003
20	Leopard	EF551002
21	Indian rhinoceros	X97336
22	White rhinoceros	Y07726
23	Black bear	DQ402478
24	Brown bear	AF303110
25	Polar bear	AF303111
26	Giant panda	EF212882
27	Rabbit	AJ001588
28	Hedgehog	X88898
29	Dormouse	AJ001562
30	Squirrel	AJ238588
31	Blue whale	X72204

where $\cos(v_1, v_2)$ is the cosine angle of vectors v_1 and v_2, and $|v_1|, |v_2|$ are the norms of vector v_1 and v_2, respectively.

Once the distance matrix exhibiting the distances among all genetic sequences considered in sequence comparison is obtained, the result of sequence comparison will be shown by the Neighbor-Joining (NJ) tree under the software of MEGA 5.10 (Tamura et al., 2011).

3. Results and discussion

In order to illustrate the usefulness of our k-mer sparse matrix model method for genetic sequence, the proposed method is applied to sequence comparison on the datasets of the mitochondrial genome sequences and 18S rRNA sequences, in which both long and short sequences are considered.

Firstly, the mitochondrial genome sequences of 31 species are applied to sequence comparison. This dataset has been analyzed in the paper of Deng et al. (2011). The descriptions of the 31 mitochondrial genome sequences are listed in Table 1, the lengths of which are from 16,338 to 17,447 base pairs (bp). The mitochondrial genetic sequence is not highly conserved for having a rapid

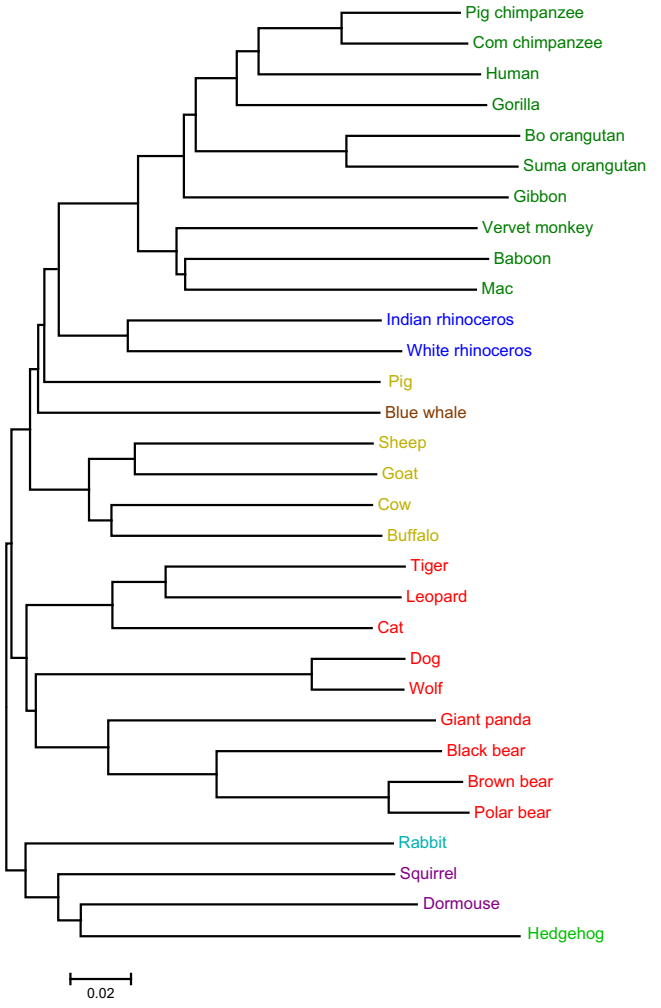

Fig. 1. The NJ tree of 31 mitochondrial genome sequences based on the 7-mer sparse matrix model. All 31 genomes are correctly clustered into eight known clusters: Primates (green), Perissodactyla (blue), Cetacea (brown), Artiodactyla (yellow), Carnivora (red), Lagomorpha (light blue), Rodentia (purple), and Erinaceomorpha (light green), similar to the results drawn by sequence alignment, standard biological taxonomy, and some published papers. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

mutation rate, so it is suitable for exploring the evolutionary relationships of different species (Yu et al., 2010; Huang et al., 2011). The NJ tree of the 31 mitochondrial genomes based on our proposed method is shown in Fig. 1 when $k=7$.

Looking at Fig. 1, all 31 genomes are correctly clustered into eight known clusters: Primates (green), Perissodactyla (blue), Cetacea (brown), Artiodactyla (yellow), Carnivora (red), Lagomorpha (light blue), Rodentia (purple), and Erinaceomorpha (light green). Because the whale was considered evolving from primitive artiodactyl, blue whale groups with artiodactyls, which are closer to rhinoceroses. This result obtained is consistent with the Euungulata Theory. Additionally, rabbit is closer to dormouse and squirrel, in that, they are all in Glires. The results shown in Fig. 1 agree well with those from standard biological taxonomy, and with results shown in some published papers (Yu et al., 2010; Huang et al., 2011; Liu et al., 2001; Raina et al., 2005; Kullberg et al., 2006). Compared with Fig. 3 of Deng et al. (2011), the accuracy of sequence comparison has been greatly improved, which can be easily seen from the relationships within the subgroups of Primates and Carnivora, respectively.

To further show the advantage of our proposed method, sequence alignment method is applied to the same dataset that we consider, using MEGA 5.10 implementation of the ClustalW algorithm. The ClustalW is a very classic sequence alignment method that calculates the best match for the selected sequences. As a contrast, the NJ tree drawn by ClustalW is shown in Fig. 2, where the species are colored the same as in Fig. 1. Although our result shown in Fig. 1 is very similar to that of ClustalW shown in Fig. 2, our result is better, as can be found from the relationship between rabbit and squirrel-dormouse.

Additionally, our proposed method is applied to the phylogenetic analysis of 40 tetrapod 18S rRNAs. The 18S rRNAs were considered odd in the estimation of phylogeny with significantly differences in higher organisms (Huelsenbeck et al., 1996). The evolutionary relationships among tetrapod have been widely discussed, and a controversial problem among tetrapod is whether birds are closer to crocodilians, or to mammals. Previous phylogenetic analysis of tetrapod 18S rRNAs supported the grouping of birds and mammals (Xia et al., 2003), whereas data from molecules, morphology and paleontology favored birds clustering with crocodilians (Hedges et al., 1990), which is more acceptable in

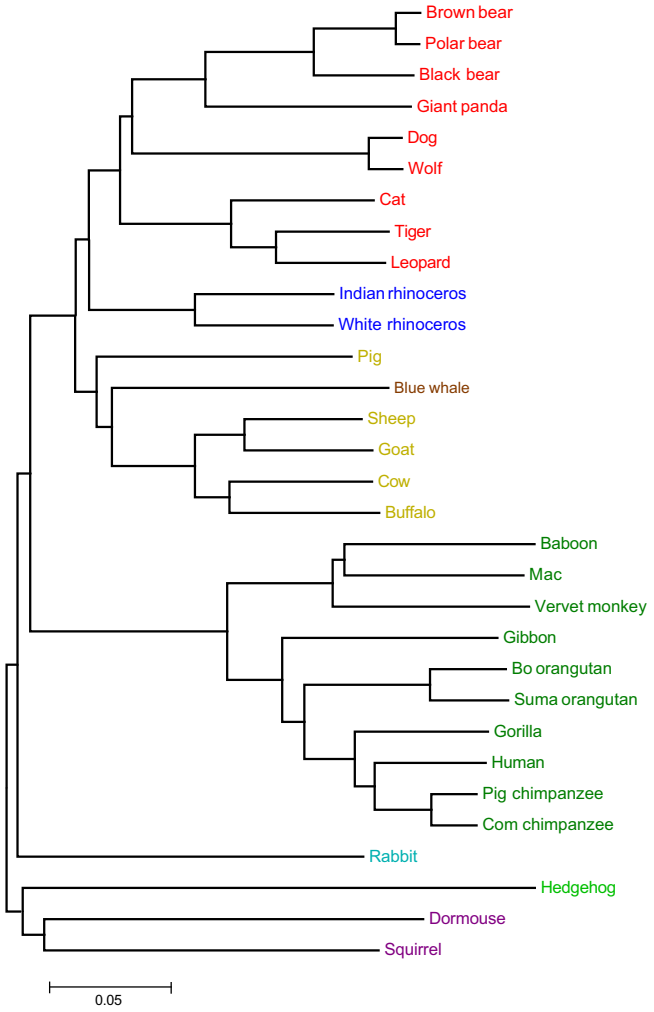

Fig. 2. The NJ tree of 31 mitochondrial genome sequences obtained by ClustalW.

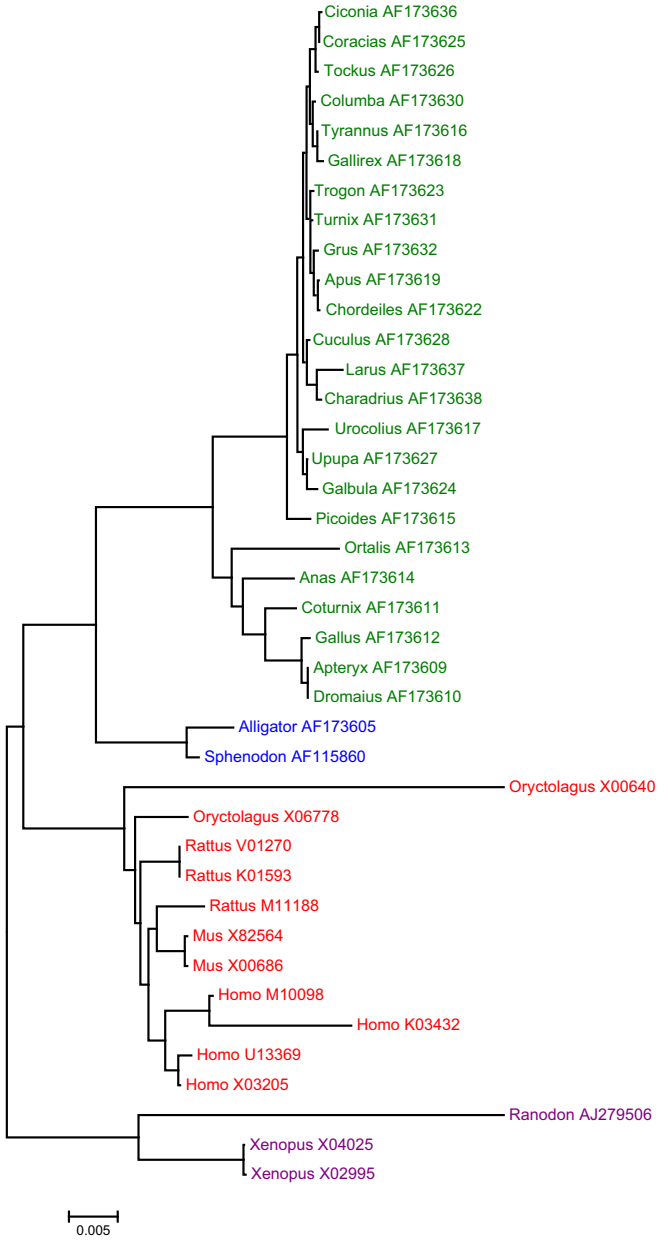

Fig. 3. The NJ tree of 40 tetrapod 18S rRNA sequences based on the 5-mer sparse matrix model. The NJ tree of 18S rRNAs contains four clades: Birds (green), Crocodilians (blue), Mammals (red) and Amphibians (purple), and the species in each clade are correctly grouped together conforming to results from traditional classification in molecules, morphology, and paleontology. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

biology. To further explore this problem, we apply our proposed method to the tetrapod 18S rRNAs dataset shown in Fig. 3 of Chan et al., (2012), which contains 40 sequences whose lengths are from 1733 to 2235 base pairs (bp), and our NJ tree is shown in Fig. 3 when $k=5$.

This NJ tree contains four clades: Birds (green), Crocodilians (blue), Mammals (red), and Amphibians (purple), and the species in each clade are correctly grouped together. This result is similar to those obtained by sequence alignment and some phylogenetic analyses (Xia et al., 2003; Hedges et al., 1990; Chan et al., 2012; Rzhetsky and Nei, 1992; Hedges, 1994; Seutin et al., 1994; Janke and Arnason, 1997; Zardoya and Meyer, 1998; Ausio et al., 1999; Dixon and Hillis, 1993). Moreover, it can be seen that birds are closer to crocodilians in Fig. 3, rather than to mammals obtained by the ClustalW shown in Fig. 4. Our result conforms to the traditional classification and evidences from molecules, morphology, and paleontology, and the results obtained by Hedges et al. (1990) and Chan et al. (2012). Compared with Fig. 3 of Chan et al. (2012), our NJ tree is much better in that not only Rattus and Mus cluster together, but also Oryctolagus are well grouped.

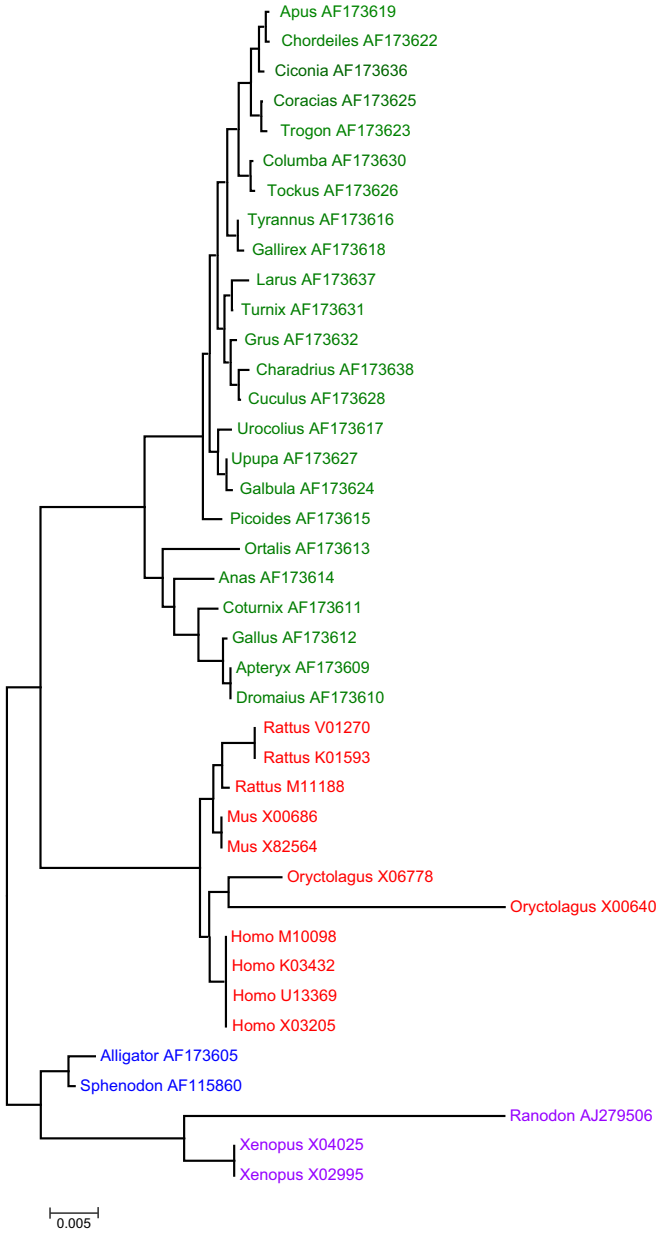

Fig. 4. The NJ tree of 40 tetrapod 18S rRNA sequences obtained by clustalW.

Therefore, based on our results from the applications to mitochondrial genome sequences and 18S rRNA sequences, it is shown that our k-mer sparse matrix model method not only depicts well the relationships of genetic sequences, but also improves the accuracy of former proposed methods. Importantly, our result shows that the birds should be close to crocodilians, rather than to mammals, which is a very valuable result for researching the evolutionary relationships of species. This result cannot be gotten by alignment-based methods. Moreover, we evaluate the optimum value k^* chosen for our k-mer sparse matrix model, and infer that the optimum value k^* is close the value of $\mathrm{floor}(\log_4 \mathrm{mean}(L))$, where L is the length set of genetic sequences considered, which is the first time to explicitly point out the optimum value of k for the k-mer model method.

4. Conclusions

In this paper, the k-mer sparse matrix model method is proposed for the comparison of genetic sequences. Based on the k-mer model of genetic sequence, the types and sites of k-mers appearing in a genetic sequence are denoted by a k-mer sparse matrix, and there exists a one-to-one relationship between a genetic sequence and its associated k-mer sparse matrix. With this representation and for a given k, a genetic sequence can be numerically characterized by a k-mer singular value vector. Moreover, we investigate and evaluate the optimum value k^* for our k-mer sparse matrix model. Our proposed method is applied in the comparison of genetic sequences, and the results obtained show that our k-mer sparse matrix model method is a very powerful method in analyzing and determining the relationships of genetic sequences.

Since user-friendly and publicly accessible web-servers represent the future direction for developing more practical and useful models, we shall make efforts to provide a web-server for the method presented in this paper. However, our k-mer sparse matrix model is still in the process of being improved, to be further optimized in the future work.

Conflict of interest

We certify that there is no conflict of interest.

Acknowledgements

We thank Prof. Hing Sun Luk for his critically reading and editing our manuscript. We also thank anonymous reviewers for hard work and good suggestions. This work is supported by Scientific Research Fund of Heilongjiang Provincial Education Department (No. 12513097).

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.jtbi.2014.08.028.

References

Ausio, J., Soley, J.T., Burger, W., Lewis, J.D., Barreda, D., Cheng, K.M., 1999. The histidine-rich protamine from ostrich and tinamou sperm: a link between reptile and bird protamines. Biochemistry 38, 180–184.

Berry, M.W., Drmac, Z., Jessup, E.R., 1999. Matrices, vector spaces, and information retrieval. SIAM Rev. 41, 335–362.

Blaisdell, B.E., 1986. A measure of the similarity of sets of sequences not requiring sequence alignment. Proc. Nat. Acad. Sci. U.S.A 83, 5155–5159.
Chan, R.H., Chan, T.H., Yeung, H.M., Wang, R.W., 2012. Composition vector method based on maximum entropy principle for sequence comparison. IEEE/ACM Trans. Comput. Biol. Bioinf. 9, 79–87.
Chen, W., Feng, P.M., Lin, H., Chou, K.C., 2013. iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition. Nucleic Acids Res. 41, e68.
Chen, W., Lei, T.Y., Jin, D.C., Lin, H., Chou, K.C., 2014. PseKNC: a flexible web-server for generating pseudo K-tuple nucleotide composition. Anal. Biochem. 456, 53–60.
Chen, W., Lin, H., Feng, P.M., Ding, C., Zuo, Y.C., Chou, K.C., 2012. iNuc-PhysChem: a sequence-based predictor for identifying nucleosomes via physicochemical properties. PLoS One 7, e47843.
Cheng, J., Cao, F., Liu, Z., 2013. AGP: a multimethods web server for alignment-free genome phylogeny. Mol. Biol. Evol. 30, 1032–1037.
Chou, K.C., 2001. Prediction of protein cellular attributes using pseudo-amino acid composition. Proteins 43, 246–255.
Chou, K.C., 2005. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. Bioinformatics 21, 10–19.
Chou, K.C., 2011. Some remarks on protein attribute prediction and pseudo amino acid composition. J. Theor. Biol. 273, 236–247.
Chou, K.C., Shen, H.B., 2010. Plant-mPLoc: a top-down strategy to augment the power for predicting plant protein subcellular localization. PLoS One 5, e11335.
Chou, K.C., Wu, Z.C., Xiao, X., 2012. iLoc-Hum: using accumulation-label scale to predict subcellular locations of human proteins with both single and multiple sites. Mol. Biosyst. 8, 629–641.
Dai, Q., Liu, X., Yao, Y., Zhao, F., 2011. Numerical characteristics of word frequencies and their application to dissimilarity measure for sequence comparison. J. Theor. Biol. 276, 174–180.
Deng, M., Yu, C., Liang, Q., He, R.L., Yau, S.S., 2011. A novel method of characterizing genetic sequences: genome space with biological distance and applications. PLoS One 6, e17293.
Dixon, M.T., Hillis, D.M., 1993. Ribosomal RNA secondary structure: compensatory mutations and implications for phylogenetic analysis. Mol. Biol. Evol. 10, 256–267.
Esmaeili, M., Mohabatkar, H., Mohsenzadeh, S., 2010. Using the concept of Chou's pseudo amino acid composition for risk type prediction of human papillomaviruses. J. Theor. Biol. 263, 203–209.
Feng, P.M., Chen, W., Lin, H., Chou, K.C., 2013. iHSP-PseRAAAC: identifying the heat shock protein families using pseudo reduced amino acid alphabet composition. Anal. Biochem. 442, 118–125.
Gao, Y., Luo, L., 2012. Genome-based phylogeny of dsDNA viruses by a novel alignment-free method. Gene 492, 309–314.
Guo, S.H., Deng, E.Z., Xu, L.Q., Ding, H., Lin, H., Chen, W., Chou, K.C., 2014. iNuc-PseKNC: a sequence-based predictor for predicting nucleosome positioning in genomes with pseudo k-tuple nucleotide composition. Bioinformatics 30, 1522–1529.
Hajisharifi, Z., Piryaiee, M., Mohammad, Beigi.M., Behbahani, M., Mohabatkar, H., 2014. Predicting anticancer peptides with Chou′ s pseudo amino acid composition and investigating their mutagenicity via Ames test. J. Theor. Biol. 341, 34–40.
Hayat, M., Khan, A., 2012. Discriminating outer membrane proteins with fuzzy K-nearest neighbor algorithms based on the general form of Chou's PseAAC. Protein Pept. Lett 19, 411–421.
Hedges, S.B., 1994. Molecular evidence for the origin of birds. Proc. Nat. Acad. Sci. U.S.A 91, 2621–2624.
Hedges, S.B., Moberg, K.D., Maxson, L.R., 1990. Tetrapod phylogeny inferred from 18S and 28S ribosomal RNA sequence and a review of the evidence for amniote relationships. Mol. Biol. Evol. 7, 607–633.
Huang, G., Zhou, H., Li, Y.F., Xu, L., 2011. Alignment-free comparison of genome sequences by a new numerical characterization. J. Theor. Biol. 281, 107–112.
Huelsenbeck, J.P., Bull, J.J., Cunningham, C.W., 1996. Combining data in phylogenetic analysis. Trends Ecol. Evol. 11, 152–158.
Janke, A., Arnason, U., 1997. The complete mitochondrial genome of Alligator mississippiensis and the separation between recent archosauria (birds and crocodiles). Mol. Biol. Evol. 14, 1266–1272.
Jun, S.R., Sims, G.E., Wu, G.A., Kim, S.H., 2010. Whole-proteome phylogeny of prokaryotes by feature frequency profiles: an alignment-free method with optimal feature resolution. Proc. Nat. Acad. Sci. U.S.A 107, 133–138.
Kantorovitz, M.R., Robinson, G.E., Sinha, S., 2007. A statistical method for alignment-free comparison of regulatory sequences. Bioinformatics 23, i249–i255.
Khosravian, M., Kazemi Faramarzi, F., Mohammad Beigi, M., Behbahani, M., Mohabatkar, H., 2013. Predicting antibacterial peptides by the concept of Chou's pseudo-amino acid composition and machine learning methods. Protein Pept. Lett. 20, 180–186.
Korf, I.F., Rose, A.B., 2009. Applying word-based algorithms: the IMEter. Methods Mol. Biol. 553, 287–301.
Kullberg, M., Nilsson, M., Arnason, U., Harley, EH., Janke, A., 2006. Housekeeping genes for phylogenetic analysis of eutherian relationships. Mol. Biol. Evol. 23, 1493–1503.
Liu, F.G., Miyamoto, M.M., Freire, N.P., Ong, P.Q., Tennant, M.R., Yong, T.S., Gugel, K.F., 2001. Molecular and morphological supertrees for eutherian (placental) mammals. Science 291, 1786–1789.
Lin, W.Z., Fang, J.A., Xiao, X., Chou, K.C., 2011. iDNA-Prot: identification of DNA binding proteins using random forest with grey model. PLoS One 6, e24756.
Mohabatkar, H., 2010. Prediction of cyclin proteins using Chous pseudo amino acid composition. Protein Pept. Lett 17, 1207–1214.
Mohabatkar, H., Mohammad Beigi, M., Abdolahi, K., Mohsenzadeh, S., 2013. Prediction of allergenic proteins by means of the concept of Chou's pseudo amino acid composition and a machine learning approach. Med. Chem. 9, 133–137.
Mohammad, Beigi.M., Behjati, M., Mohabatkar, H., 2011. Prediction of metalloproteinase family based on the concept of Chou's pseudo amino acid composition using a machine learning approach. J. Struct. Funct. Genomics 12, 191–197.
Mondal, S., Pai, P.P., 2014. Chou' s pseudo amino acid composition improves sequence-based antifreeze protein prediction. J. Theor. Biol. 13, 30–35.
Nanni, L., Lumini, A., 2008. Genetic programming for creating Chou's pseudo amino acid based features for submitochondria localization. Amino Acids 34, 653–660.
Nanni, L., Lumini, A., Gupta, D., Garg, A., 2012. Identifying bacterial virulent proteins by fusing a set of classifiers based on variants of Chou's pseudo amino acid composition and on evolutionary information. IEEE/ACM Trans. Comput. Biol. Bioinf 9, 467–475.
Qi, J., Wang, B., Hao, B.L., 2004. Whole proteome prokaryote phylogeny without sequence alignment: a k-string comparison approach. J. Mol. Evol. 58, 1–11.
Qiu, W.R., Xiao, X., Chou, K.C., 2014. iRSpot-TNCPseAAC: identify recombination spots with trinucleotide composition and pseudo amino acid components. Int. J. Mol. Sci. 15, 1746–1766.
Raina, S.Z., Faith, J.J., Disotell, T.R., Seligmann, H., Stewart, C.B., Pollock, D.D., 2005. Evolution of base-substitution gradients in primate mitochondrial genomes. Genome Res. 15, 665–673.
Rzhetsky, A., Nei, M., 1992. A simple method for estimating and testing minimum-evolution tree. Mol. Biol. Evol. 9, 945–967.
Seutin, G., Lang, B.F., Mindell, D.P., Morais, R., 1994. Evolution of the WANCY region in amniote mitochondrial DNA. Mol. Biol. Evol. 11, 329–340.
Sims, G.E., Jun, S.R., Wu, G.A., Kim, S.H., 2009a. Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. Proc. Nat. Acad. Sci. U.S.A 106, 2677–2682.
Sims, G.E., Jun, S.R., Wu, G.A., Kim, S.H., 2009b. Whole-genome phylogeny of mammals: evolutionary information in genic and non-genic regions. Proc. Nat. Acad. Sci. U.S.A 106, 17077–17082.
Stuart, G.W., Berry, M.W., 2004. An SVD-based comparison of nine whole eukaryotic genomes supports a coelomate rather than ecdysozoan linkage. BMC Bioinf. 5, 204.
Stuart, G.W., Moffett, K., Leader, J.J., 2002. A comprehensive vertebrate phylogeny using vector representations of protein sequences from whole genomes. Mol. Biol. Evol. 19, 554–562.
Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., Kumar, S., 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol. Biol. Evol. 28, 2731–2739.
Wen, J., Zhang, Y.Y., 2009. A 2D graphical representation of protein sequence and its numerical characterization. Chem. Phys. Lett. 476, 281–286.
Wen, J., Chan, R.H, Yau, S.C., He, R.L., Yau, S.T., 2014. *K*-mer natural vector and its application to the phylogenetic analysis of genetic sequences. Gene 546, 25–34.
Wu, T.J., Burke, J.P., Davison, D.B., 1997. A measure of DNA sequence dissimilarity based on Mahalanobis distance between frequencies of words. Biometrics 53, 1431–1439.
Wu, T.J., Hsieh, Y.C., Li, L.A., 2001. Statistical measures of DNA dissimilarity under Markov chain models of base composition. Biometrics 57, 441–448.
Wu, T.J., Huang, Y.H., Li, L.A., 2005. Optimal word sizes for dissimilarity measures and estimation of the degree of dissimilarity between DNA sequences. Bioinformatics 21, 4125–4132.
Xia, X., Xie, Z., Kjer, K.M., 2003. 18S ribosomal RNA and tetrapod phylogeny. Syst. Biol. 52, 283–295.
Xiao, X., Wu, Z.C., Chou, K.C., 2011. iLoc-Virus: a multi-label learning classifier for identifying the subcellular localization of virus proteins with both single and multiple sites. J. Theor. Biol. 284, 42–51.
Yang, X.W., Wang, T.M., 2013. A novel statistical measure for sequence comparison on the basis of k-word counts. J. Theor. Biol. 318, 91–100.
Xu, Y., Ding, J., Wu, L.Y., Chou, K.C., 2013. iSNO-PseAAC: predict cysteine S-nitrosylation sites in proteins by incorporating position specific amino acid propensity into pseudo amino acid composition. PLoS One 8, e55844.
Yang, L., Zhang, X., Zhu, H., 2013. Alignment free comparison: k word voting model and its applications. J. Theor. Biol. 335, 276–282.
Yu, C., Liang, Q., Yin, C., He, R.L., Yau, S.S., 2010. A Novel Construction of Genome Space with Biological Geometry. DNA Res 17, 155–168.
Yu, H.J., 2013. Segmented *K*-mer and its application on similarity analysis of mitochondrial genome sequences. Gene 518, 419–424.
Yu, H.J., Huang, D.S., 2012. Novel 20-D descriptors of protein sequence and its applications in similarity analysis. Chem. Phys. Lett. 531, 261–266.
Zardoya, R., Meyer, A., 1998. Complete mitochondrial genome suggests diapsid affinities of turtles. Proc. Nat. Acad. Sci. U.S.A 95, 14226–14231.
Zou, D., He, Z., He, J., Xia, Y., 2011. Supersecondary structure prediction using Chou's pseudo amino acid composition. J. Comput. Chem. 32, 271–278.