Gene 529 (2013) 250–256

Contents lists available at ScienceDirect

Gene

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

Protein sequence comparison based on *K*-string dictionary

Chenglong Yu [a], Rong L. He [b], Stephen S.-T. Yau [c,*]

[a] Department of Mathematics, Statistics and Computer Science, University of Illinois at Chicago, IL 60607-7045,USA
[b] Department of Biological Sciences, Chicago State University, Chicago, IL,USA
[c] Department of Mathematical Sciences, Tsinghua University, Beijing, PR China

ARTICLE INFO

Article history:
Accepted 25 July 2013
Available online 9 August 2013

Keywords:
K-string
Sequence comparison
Frequency vector
Cardinality
Singular Value Decomposition

ABSTRACT

The current *K*-string-based protein sequence comparisons require large amounts of computer memory because the dimension of the protein vector representation grows exponentially with *K*. In this paper, we propose a novel concept, the "*K*-string dictionary", to solve this high-dimensional problem. It allows us to use a much lower dimensional *K*-string-based frequency or probability vector to represent a protein, and thus significantly reduce the computer memory requirements for their implementation. Furthermore, based on this new concept, we use Singular Value Decomposition to analyze real protein datasets, and the improved protein vector representation allows us to obtain accurate gene trees.

1. Introduction

With the development of biotechnology, more and more biological sequences have been acquired. The discovery of new protein sequences is accelerating, but many of these proteins show similarity to existing amino acid sequences. Sequence comparison problems arise when detecting the similarity of proteins, and explaining their phylogenetic relations as well as when handling the huge amount of data. Existing methods for sequence comparison can be classified into alignment-based methods and alignment-free methods. Alignment-based methods use dynamic programming, a regression technique that finds an optimal alignment by assigning scores to different possible alignments and picking the alignment with the highest score (Gotoh, 1982; Needleman and Wunsch, 1970; Smith and Waterman, 1981). However, the search for optimal solutions using sequence alignment turns out to be computationally difficult with large biological databases, especially when comparing three or more biological sequences at a time, i.e., multiple sequence alignment. Therefore, alignment-free approaches have been developed to overcome the critical limitations of alignment-based methods.

The recent reviews (Davies et al., 2008; Vinga and Almeida, 2003) on published methods of alignment-free sequence comparison report several concepts of distance measures, such as Markov chain models and Kullback–Leibler discrepancy (Wu et al., 2001), chaos theory (Almeida et al., 2001), Kolmogorov complexity (Li et al., 2001), decision tree induction algorithm (Huang et al., 2004), graphical representation (Liao and Wang, 2004; Randic et al., 2003; Yau et al., 2003), probabilistic measure (Pham and Zuegg, 2004; Yu et al., 2011a,b), and pseudo amino acid composition (Chou, 2011; Chou and Shen, 2009). Furthermore, sequence vector representation approaches without alignment are also prevalent, such as feature vector (Carr et al., 2010; Liu et al., 2006), moment vector (Yau et al., 2008; Yu et al., 2010, 2011a,b), and natural vector (Deng et al., 2011; Yu et al., 2013). Among all existing alignment-free methods, the *K*-string-based methods (Chu et al., 2004; Gao and Qi, 2007; Lu et al., 2008; Qi et al., 2004; Takahashi et al., 2009) have received substantial attention. Basically, the first step of these methods is, for a fixed integer *K*, to count the number of overlapping *K*-peptides in one protein sequence, and form a frequency or probability vector of dimension 20^K. Then using some probabilistic or optimization models these vectors are converted into more complicated composition vectors (Chan et al., 2012), but the dimension of the vectors remains unchanged in this process. Finally, the distance between two composition vectors is used to compute the distance between two taxa, and once the distances among all taxa are obtained, the phylogenetic trees can be reconstructed. These methods are able to provide good phylogenetic tree topologies for DNA or proteins; however, because large values needed to be chosen (see the discussion in Section 2), the resulting high memory usage becomes a disadvantage.

In this paper, we provide a novel concept, the "*K*-string dictionary", to solve this problem. It allows us to use a much lower dimensional frequency or probability vector to represent a protein, and thus significantly reduce the memory requirements for their implementation. Furthermore, after obtaining the lower dimensional frequency vectors, we use Singular Value Decomposition (SVD) to get an improved protein vector representation which allows us to obtain accurate gene trees. We have analyzed 290 proteins from 3 families and 50 beta-globin

Abbreviations: SVD, Singular Value Decomposition; MSA, multiple sequence alignment; ND1, NADH dehydrogenase 1.

* Corresponding author. Department of Mathematical Sciences, Tsinghua University, Beijing 100084, PR China. Tel.: +86 10 62787874; fax: +86 10 62798033.
E-mail address: yau@uic.edu (S.S.-T. Yau).

0378-1119/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.gene.2013.07.092

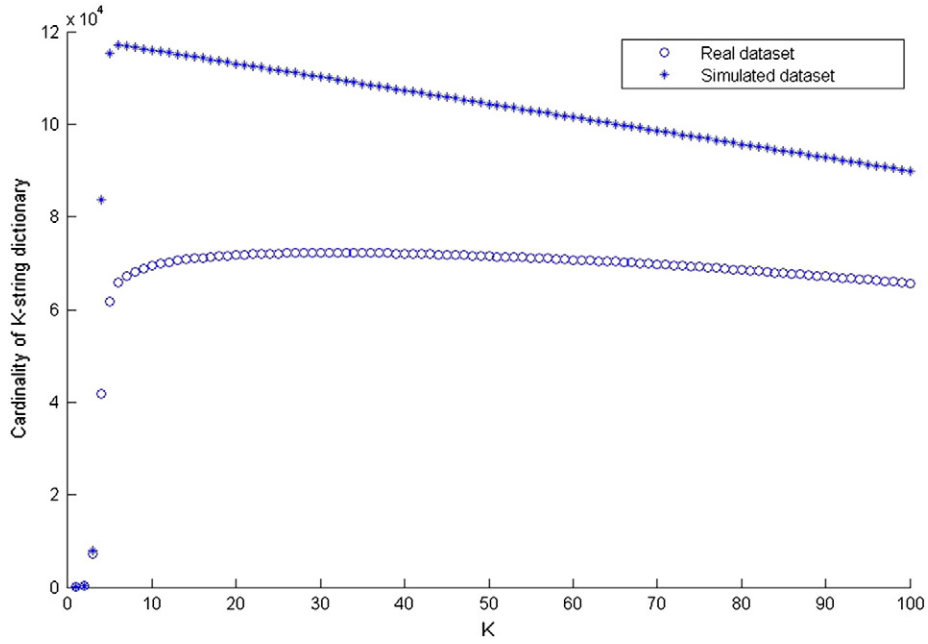

Fig. 1. The cardinalities of *K*-string dictionary of real and simulated datasets including 290 proteins.

proteins from different animal species using this method, and found it to be a powerful classification tool for proteins.

2. Materials and methods

2.1. Background on K-string frequency or probability vector

Given a protein sequence of length L, the frequency of appearances of a K-string $\alpha = a_1a_2, \ldots a_K$ in this sequence is defined as $f(\alpha)$, where α_i is an amino acid single-letter symbol. This frequency divided by the total number $(L - K + 1)$ of K-strings in the given protein sequence is defined as the probability $p(\alpha)$ of appearance of the K-string α in the sequence: $p(\alpha) = \frac{f(\alpha)}{L-K+1}$. For example, given a protein sequence (AMFAMCAMFS), $f(\alpha) = 2$ for 3-string $\alpha =$ (AMF), and $p(\alpha) = \frac{2}{10-3+1} = 0.25$.

Table 1
The cardinalities of *K*-string dictionary of real and simulated dataset.

	Cardinality	
K value	Real dataset	Simulated dataset
1	20	20
2	400	400
3	7186	8000
4	41703	83601
5	61792	115394
6	65733	117083
7	67214	116892
8	68182	116604
9	68898	116314
10	69450	116024
11	69895	115734
12	70255	115444
13	70551	115154
14	70804	114864
15	71012	114574
16	71188	114284
17	71343	113994
18	71482	113704
19	71607	113414
20	71720	113124

There are a total of $N = 20^K$ possible types of such K-strings for protein sequences. Thus the K-string frequency vector of one protein sequence is defined as $(f(\alpha_1), f(\alpha_2), \ldots, f(\alpha_N))$, and the corresponding K-string probability vector of one protein sequence is defined as $(p(\alpha_1), p(\alpha_2), \ldots, p(\alpha_N))$.

Many current alignment-free works are based on the K-string frequency or probability vectors as we mentioned in Section 1. However, the choice of suitable K has always been an important concern. The main problem is that the dimension of these vectors can quickly become large. For example, the dimension of the protein K-string frequency or probability vector for $K = 6$ is $20^6 = 64{,}000{,}000$. Trying to work with vectors of such a large dimension will exceed the memory limits of ordinary personal computers. Thus, when using these vectors, we cannot evaluate the results for larger K. To overcome this disadvantage, we propose a novel concept "K-string dictionary" to solve this problem.

2.2. K-string dictionary

The K-string dictionary of a group of protein sequences is the set of all K-strings existing in these sequences. Note that a set is a collection of distinct objects, so we only record repeated K-strings once in the dictionary. For example, given a group of two protein sequences (AMTHGS) and (MTHAKW), the 3-string dictionary for this group is the set {AMT, MTH, THG, HGS, THA, HAK, AKW}. The key point is that the cardinality of a K-string dictionary is far less than 20^K. This will significantly reduce the memory requirements for computer calculations.

For example, titin is currently the largest known protein; its human variant (GenBank No.: NP_001243779) consists of 34,350 amino acids (Minajeva et al., 2001). For example, we take $K = 10$, then titin has $34{,}350 - 10 + 1 = 34{,}341$ K-strings. Assume that we are dealing with 1000 big proteins like titin's size, and all 10-strings of them are totally different, then the cardinality of the 10-string dictionary of this group is $34{,}314 \times 1000 = 3.4341 \times 10^7$. However, this number is still far less than $20^{10} = 1.024 \times 10^{13}$.

2.3. The cardinality of K-string dictionary

Given a group of protein sequences, for different K, we have different K-string dictionaries. We will use the real and simulated protein

Fig. 2. The neighbor-joining phylogenetic tree of 290 proteins from 3 families based on 3-string dictionary.

sequence datasets to investigate the cardinalities of different *K*-string dictionaries.

The real dataset consists of the 290 proteins belonging to three families (PF03296, PF06924, and PF09455) in the Pfam database (Bateman et al., 2002). Pfam is a database of protein families that includes their annotations and multiple sequence alignments generated using hidden Markov models. Furthermore, according to SCOP (Lo Conte et al., 2000), a largely manual classification database of protein structural domains, all of these 290 proteins are multidomain proteins. The PF03296 family belongs to Poly (A) polymerase catalytic subunit-like Fold (SCOP). They have three domains; D1: all-alpha, contains HhH motif; D2: alpha + beta of nucleotidyltransferase fold (scop_cf 81302); D3: alpha + beta; beta(3)-alpha-beta(3)-alpha(2). The PF06924 family belongs to Api92-like Fold (SCOP). They have two domains; D1: alpha + beta with similarity to ferredoxin-like fold; D2: 6 helices, bundle, one buried central helix, inserted into D1. The PF09455 family belongs to SSO1389-like Fold (SCOP). They have two domains; D1: alpha/beta, central parallel beta-sheet of 6 strands, order 321456, Rossmann-like; D2: alpha + beta, cluster of helices and a small 4-stranded beta-sheet. PF03296 family has 53 proteins; PF06924 family has 83 proteins; PF09455 family has 154 proteins. For details of this dataset, please see Supplementary materials.

Based on this real dataset, we generate one simulated dataset. It also includes 290 protein sequences which have the same lengths as the real sequences in the above real dataset, but each sequence has the amino acid content of equal probability.

In Fig. 1, we show the cardinalities of *K*-string dictionary of these two datasets. The cardinality is first increasing then decreasing with the increase of *K*. It is due to the fact that, when *K* becomes larger and approaching the length of the sequence, the number of *K*-strings in the sequence becomes smaller. Thus, we can obtain the maximum cardinality value for some *K* value. We also find that the cardinalities of the *K*-string dictionary of the simulated dataset are always larger than those of real dataset when $K \geq 3$, as shown in Table 1. Furthermore, after the cardinality of the *K*-string dictionary of the simulated dataset reaches the maximum value, it is linearly decreasing. The reason for these is because we generate the simulated sequences by assuming equal probability for each amino acid. In this case, the probability that two *K*-strings agree is $1/20^K$. Thus, with the increase of *K*, this probability becomes very small. This implies that all *K*-strings are different in this simulated dataset when *K* is large. So, suppose there is a group of *n* simulated sequences with amino acid content of equal probability, and each sequence has the length L_i ($i = 1, 2, \ldots, n$). Then, when *K* is large, the cardinality of the *K*-string dictionary for this group is $\sum_{i=1}^{n}(L_i - K + 1)$, and also, when *K* increases by 1, the cardinality decreases by *n*. This explains that why the cardinality is linearly decreasing after the peak. In Table 1, we can see that when *K* is larger than 8, the cardinality of the simulated dataset decreases by 290 with the increase of *K*. Furthermore, the cardinality of both the real and simulated datasets is clearly far less than 20^K.

2.4. New K-string frequency or probability vector based on K-string dictionary

After obtaining the *K*-string dictionary, we can redefine the *K*-string frequency or probability vector. Given a group of protein sequences, let *D* be the *K*-string dictionary of this group: $D = \{d_1, d_2, \ldots, d_c\}$, where d_i is the *K*-string in *D* and *c* is the cardinality of *D*.

For one sequence in this group, the frequency of appearances of a *K*-string d_i in this sequence is defined as $f(d_i)$, thus the new frequency vector of the sequence is defined as $(f(d_1), f(d_2), \ldots, f(d_c))$. Clearly, if the length of this sequence is *L*, then $L - K + 1 = \sum_{i=1}^{c} f(d_i)$. Thus the corresponding *K*-string probability vector is $(p(d_1), p(d_2), \ldots, p(d_c)) = \left(\frac{f(d_1)}{\sum_{i=1}^{c} f(d_i)}, \frac{f(d_2)}{\sum_{i=1}^{c} f(d_i)}, \ldots, \frac{f(d_c)}{\sum_{i=1}^{c} f(d_i)} \right)$.

2.5. Distance measurement

Given two vectors, there are many different distances to measure their similarity/dissimilarity (Vinga and Almeida, 2003). An angle-based distance is widely used when dealing with DNA sequences. Let

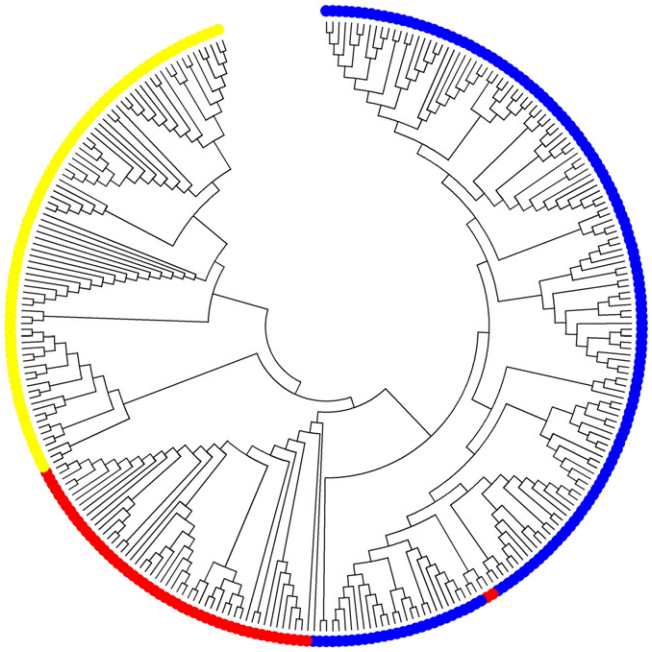

Fig. 3. The neighbor-joining phylogenetic tree of 290 proteins from 3 families based on 4-string dictionary.

Table 2
50 beta-globin sequences of animal species.

Animal names	Accession number
Human	AAA16334.1
Goshawk	P08851.1
Lesser panda	P18982.1
Giant panda	P18983.2
Sheep	P02075.2
Duck	P02114.2
Mallard	P02115.1
Goose	P02117.1
Rat	CAA33114.1
Penguin	P80216.1
Swift	P15165.1
Coyote	P60525.1
Catfish	O13163.2
Bison	P09422.1
Swan	P68945.1
Buffalo	P67820.1
Dog	P60524.1
Chimpanzee	P68873.2
Dolphin	P18990.1
Goldfish	P02140.1
Polar bear	P68011.1
Rhinoceros	P09907.1
Chicken	P02112.2
Wolf	P60526.1
Turtle	P13274.1
Pigeon	P11342.1
Black bear	P68012.1
Asiatic elephant	P02084.1
African elephant	P02085.1
Tortoise	P83123.3
Grivet	P02028.1
Gorilla	P02024.2
Shark	P02143.1
Hippopotamus	P19016.1
Horse	P02062.1
Gibbon	P02025.1
Whale	P18984.1
Bat	P24660.1
Red fox	P21201.1
Marmot	P08853.1
Salmon	Q91473.3
Sparrow	P07406.1
Pheasant	P02113.1
Flamingo	P02121.1
Pig	P02067.3
Dragonfish	ADD73488.1
Parakeet	P21668.1
Zebra	P67824.1
Cod	O13077.2
Langur	P02032.1

$\alpha = (p_1,p_2, \ldots,p_n)$ and $\beta = (q_1,q_2, \ldots,q_n)$ be two vectors, the cosine of the angle between vectors α and β is defined as

$$\cos(\alpha,\beta) = \frac{\sum_{i=1}^{n} p_i \times q_i}{\sqrt{\sum_{i=1}^{n} {p_i}^2} \times \sqrt{\sum_{i=1}^{n} {q_i}^2}}.$$

This cosine distance is not sensitive to repetitions for *K*-string based methods. For example, if a sequence X is compared with its double repetition XX, then the corresponding two vectors of the *K*-string counts will basically have the same direction in the *K*-string dictionary space. Thus the angle distance between these two vectors is roughly equal to zero. This property is of fundamental value because it automatically filter repetitions, and thus this distance is robust with respect to duplication mutation in genome or protein. Here we adopt a distance measurement $D(\alpha,\beta) = \frac{1-\cos(\alpha,\beta)}{2}$ because it is widely used and achieved a great success in the phylogenetic analysis of whole genomes of bacteria, viruses, and vertebrates (Chan et al., 2010).

2.6. SVD-based protein sequence representation

We can construct the *K*-string frequency matrix M of a group of n proteins. In the matrix, each protein is represented by a column of a new *K*-string frequency vector based on the *K*-string dictionary of this group of n proteins. Suppose that the cardinality of this *K*-string dictionary is c, and then the matrix M is c by n:

$$M = \begin{bmatrix} f_{11} & f_{12} & \cdots & f_{1n} \\ f_{21} & f_{22} & \cdots & f_{2n} \\ & \cdots & & \\ f_{c1} & f_{c2} & \cdots & f_{cn} \end{bmatrix}.$$

Compared to the original frequency matrix (20^K by n), M has much smaller size. This provides an easy tool for describing proteins and allows available application of numerical linear algebra tools.

SVD, a matrix factorization method, has been applied to improve the protein frequency vectors (Stuart et al, 2002a,b). M is decomposed into three separate matrices U, Σ, and V using SVD, that is,

$$M = U * \Sigma * V^T,$$

where U is the $c \times c$ orthogonal matrix having the left singular vectors of M as its column, V is the $n \times n$ orthogonal matrix having the right singular vectors of M as its column, and Σ is the $c \times n$ diagonal matrix having the singular values $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_{\min(c,n)}$ of M in order along its diagonal. The rank r of the matrix M is equal to the number of nonzero singular values. Then the Frobenius norm of M is defined as

$$\|M\|_F = \sqrt{\sum_{j=1}^{r} \sigma_j^2}.$$

The Eckart–Young theorem (Eckart and Young, 1936) states that the distance between M and its rank-m approximations ($m \leq r$) is minimized by the approximation M_m. Here

$$M_m = U_m \Sigma_m {V_m}^T,$$

where U_m is the $c \times m$ matrix whose columns are the first m columns of U, V_m is the $n \times m$ matrix whose columns are the first m columns of V, and Σ_m is the $m \times m$ diagonal matrix whose diagonal elements are the m largest singular values of M. The theorem further shows how the norm of that distance is related to singular values of M:

$$\|M - M_m\|_F = \min_{rank(X) \leq m} \|M - X\| = \sqrt{\sigma_{m+1}^2 + \ldots + \sigma_r^2}.$$

This low-rank matrix approximation can improve the relative accuracy of protein vectors by discarding a substantial fraction of the noise (including homoplasy) in the data (Stuart et al, 2002a). If $\sigma_1, \ldots, \sigma_r$ are the positive singular values of M, then by using Frobenius norm, the singular vectors associated with any particular singular value (e.g., σ_j) accounts for the fraction $\sqrt{\frac{\sigma_j^2}{\sigma_1^2+\sigma_2^2+\ldots+\sigma_r^2}}$ of the data. So, choosing the m largest values ($m < r$) explains the fraction $\sqrt{\frac{\sigma_1^2+\sigma_2^2+\ldots+\sigma_m^2}{\sigma_1^2+\sigma_2^2+\ldots+\sigma_r^2}}$ of the data, and it also allows the approximation of the matrix from the first m singular triplets:

$$M_m = U_m \Sigma_m {V_m}^T.$$

Determining the number m of ranked singular values that best serve to separate signal from noise within the data set is challenging (Berry et al., 1999). In this study, we choose the minimum m such that $\sqrt{\frac{\sigma_1^2+\sigma_2^2+\ldots+\sigma_m^2}{\sigma_1^2+\sigma_2^2+\ldots+\sigma_r^2}} \geq 95\%$, that is, we consider less than 5% as a reasonably small change to the initial matrix. Then the columns of M_m

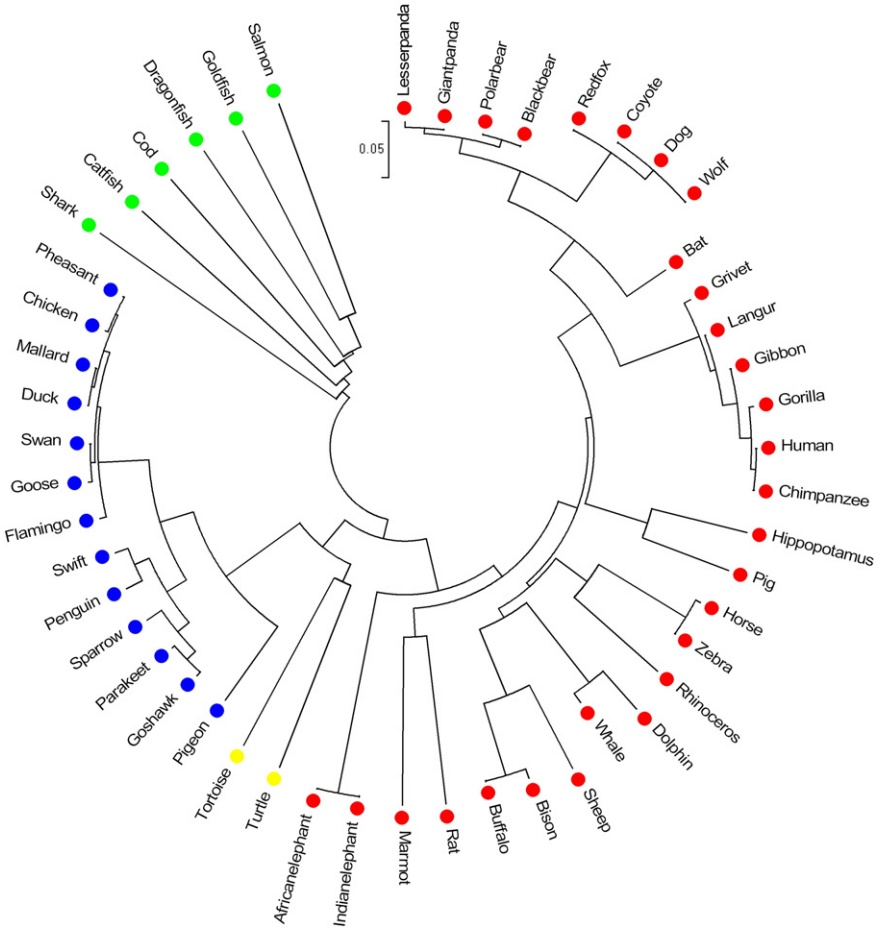

Fig. 4. The neighbor-joining phylogenetic tree of 50 beta-globin proteins from different animal species. The taxa of mammal proteins are marked by red color, the taxa of reptile proteins are marked by yellow color, the taxa of avian proteins are marked by blue color, and the taxa of fish proteins are marked by green color.

give us the improved vectors, which can be used to measure the similarity/dissimilarity of the original proteins.

3. Results

To test that the new improved vectors obtained in this way truly incorporates the classification analysis of proteins, we apply it to the real protein dataset (290 proteins) mentioned in Section 2.3. Firstly, we try $K = 3$, then the cardinality of 3-string dictionary of this dataset is 7186. So, the corresponding 3-string frequency matrix M is 7186 by 290. By using SVD mentioned in Section 2.6, we get the improved protein representation vectors. The distance matrix for the group of proteins is constructed from all pairwise vectors. We use neighbor-joining algorithm (Saitou and Nei, 1987) of MEGA 5.0 software (Tamura et al., 2011) to construct the phylogenetic tree based on the distance matrix, as shown in Fig. 2. The taxa of family PF03296 are marked by red color, the taxa of family PF06924 are marked by yellow color, and the taxa of family PF09455 are marked by blue color. We can see that family PF03296 and family PF09455 are mixed together. Then we try $K = 4$, the cardinality of 4-string dictionary of this dataset is 41,703, far less than $20^4 = 160{,}000$. The corresponding 4-string frequency matrix M is 41,703 by 290. Similarly, we get the phylogenetic tree as shown in Fig. 3. We can see that the classification result is much improved; only two taxa of family PF03296 are put into family PF09455. This illustrates that our new method can give very high classification accuracy for proteins. In Supplementary materials, we also give the traditional rectangular phylogenetic trees of Figs. 2 and 3 with more details (see Supplementary Fig. 1 and Fig. 2).

Fifty beta-globin sequences of different species (Yau et al., 2008) were extracted from GenBank as shown in Table 2. As we discussed in Section 2, the original frequency vector for $K = 6$ has dimension of $20^6 = 64{,}000{,}000$, which exceeded the memory limits of ordinary PC computers. Here we use the new frequency vector based on the 6-string dictionary. The cardinality of 6-string dictionary of these 50 beta-globins is 2051 ($\ll$64,000,000). Thus, the corresponding 6-string

Table 3
19 NADH dehydrogenase 1 protein sequences of mammal species.

Animal names	Accession number
Gibbon	NC_002082.1
Gorilla	NC_011120.1
Human	NC_012920.1
Chimp	NC_001643.1
Pygmy Chimp	NC_001644.1
Sumatran Orang	NC_002083.1
Bornean Orang	NC_001646.1
Hedgehog	NC_002080.2
Rat	AC_000022.2
Mouse	NC_005089.1
Rhino	NC_001779.1
Donkey	NC_001788.1
Horse	NC_001640.1
Cow	NC_006853.1
Baleen whale	NC_001601.1
Fin whale	NC_001321.1
Cat	NC_001700.1
Gray seal	NC_001602.1
Harbor seal	NC_001325.1

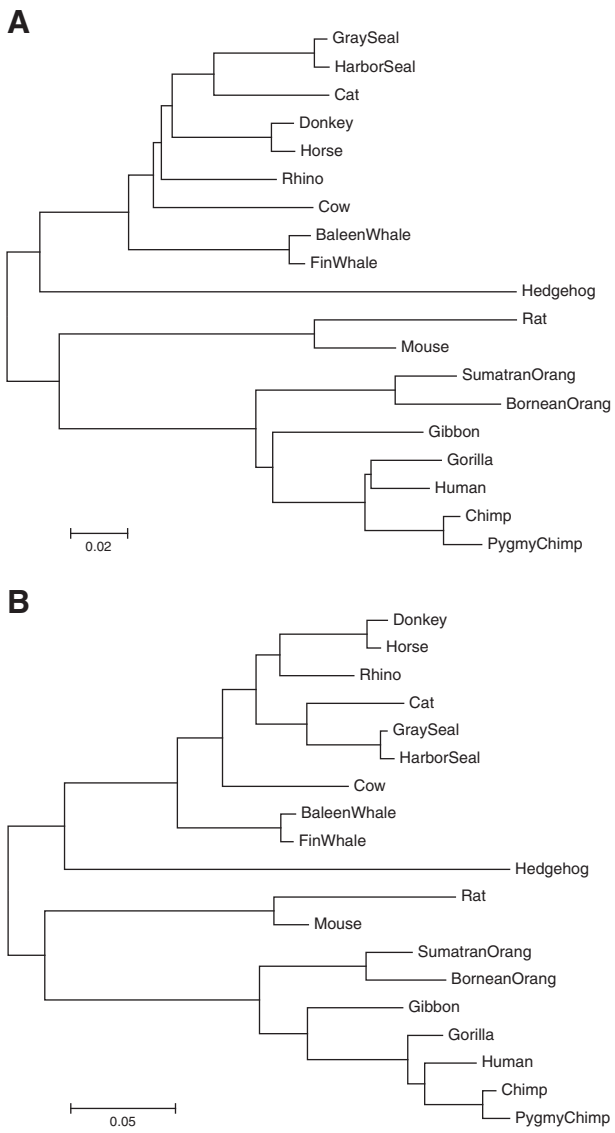

Fig. 5. (A) The neighbor-joining tree for the 19 mammalian species based on multiple sequence alignment; (B) the maximum parsimony tree for the 19 mammalian species based on multiple sequence alignment.

frequency matrix M is 2051 by 50. By using SVD, the minimum m such that $\sqrt{\frac{\sigma_1^2+\sigma_2^2+\ldots+\sigma_m^2}{\sigma_1^2+\sigma_2^2+\ldots+\sigma_r^2}} \geq 95\%$ is 23. So, the columns of $M_{23} = U_{23}\Sigma_{23}V_{23}{}^T$ give us the improved protein representation vectors. In Fig. 4, we show the phylogenetic tree of these 50 beta-globins, which is also reconstructed by neighbor-joining program of MEGA 5.0 software. We note that these 50 beta-globins are clearly separated into four clusters (mammal, avian, fish, and reptile). Furthermore, the distances between beta-globin sequences from several primate species (human, grivet, gorilla, langur, gibbon, and chimpanzee) are very small, and they form a subcluster in the resulting protein tree.

4. Discussion

In order to show the feasibility and efficiency of our approach, we compare our method with multiple sequence alignment (MSA) and other alignment-free tools. We use a new dataset including 19 NADH dehydrogenase 1 (ND1) protein sequences of different species to test the classification analysis results. The ND1 protein data plays an important role in phylogenetic classification of mammals (Cao et al., 1998). In Table 3, we gave the details of this dataset.

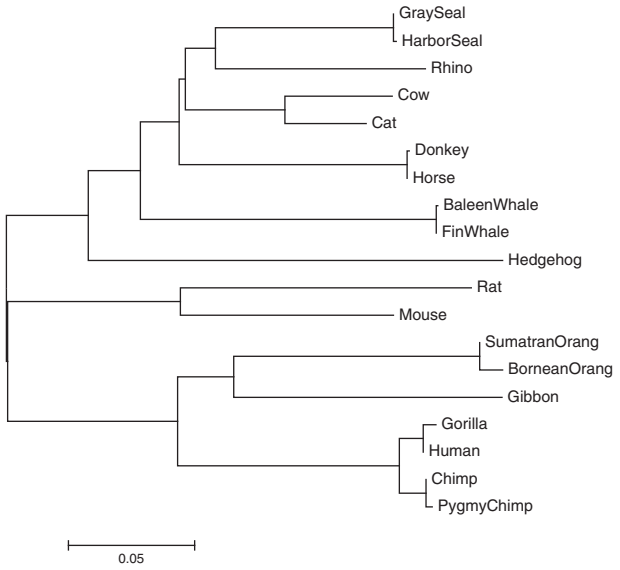

Fig. 6. The phylogenetic tree for the 19 mammalian species based on our new method.

4.1. Comparison with multiple sequence alignment (MSA) method

We use the existing alignment tool ClustalW to make the multiple sequence alignment for these 19 protein sequences. We choose the BLOSUM30 amino acid substitution matrix in this process. Figs. 5(A and B) show two phylogenetic trees based on the alignment result with neighbor-joining and maximum parsimony methods. Both trees are reconstructed by MEGA 5.0 software. On the other hand, we still use our new frequency vector based on 6-string dictionary. The cardinality of 6-string dictionary of these 19 protein sequences is 1849. Thus, the corresponding 6-string frequency matrix M is 1849 by 19. By using SVD, the minimum m such that $\sqrt{\frac{\sigma_1^2+\sigma_2^2+\ldots+\sigma_m^2}{\sigma_1^2+\sigma_2^2+\ldots+\sigma_r^2}} \geq 95\%$ is 11. So, the columns of $M_{11} = U_{11}\Sigma_{11}V_{11}{}^T$ give us the improved protein representation vectors. Here we still adopt the distance measurement $D(\alpha,\beta) = \frac{1-\cos(\alpha,\beta)}{2}$ to get the distance matrix for these vectors. In Fig. 6, we show the phylogenetic tree of these 19 sequences, which is reconstructed by neighbor-joining program of MEGA 5.0 software. We can see that our method obtains very similar results as MSA does. For example, the distances between sequences from several primate species (Sumatran Orang, Bornean Orang, Gibbon, Gorilla, Human, Chimp, and Pygmy Chimp) are very small, and they form a subcluster in the resulting protein tree.

4.2. Comparison with another alignment-free method

In 2003 Otu and Sayood (Otu and Sayood, 2003) developed an alignment-free sequence distance measure for phylogenetic tree construction based on the relative information between the sequences using Lempel–Ziv complexity. In Fig. 7, we show the neighbor-joining phylogenetic tree of these 19 sequences by using Otu and Sayood's distance method. We see that this method can also get similar results with our approach and MSA methods. Thus, our method brings a novel direction to comparative proteomic analysis at the sequence level. The novel concept "K-string dictionary" allows us to use a much lower dimensional frequency or probability vector to represent a protein, and thus save large amounts of computer memory space. With this approach, we can get a much smaller frequency matrix. This provides an easy and precise tool for describing proteins and makes it possible to use existing numerical linear algebra tools.

In this study we represent a protein sequence as a low dimensional numerical vector, but we do not consider any amino acid physicochemical properties in the vector. For example, the amino acid hydrophobicity

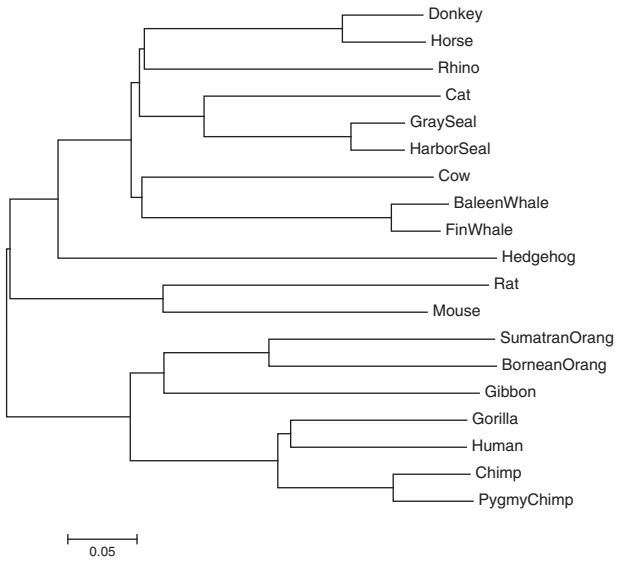

Fig. 7. The phylogenetic tree for the 19 mammalian species based on Otu and Sayood's distance method.

plays an important role in protein classification (Fauchere and Pliska, 1983). Thus further studies will be needed to combine some amino acid properties with the frequency of *K*-string in the vector. Furthermore, we adopt SVD to improve the new low dimensional frequency vector. The computational complexity for SVD is usually high. Thus, other numerical methods in matrix theory deserve further consideration and investigation in the future work.

5. Conclusion

In this paper, we propose a novel concept, the "*K*-string dictionary", to solve the high dimensional vector problem in *K*-string-based protein sequence comparisons. It allows us to use a much lower dimensional frequency or probability vector to represent a protein, and thus significantly reduce the computer memory requirements for their implementation. By using this approach, we can get much a smaller frequency matrix. This provides an easy and precise tool for describing proteins and makes it possible to use existing numerical linear algebra tools. The computer code used to prepare this paper is available from the author upon request.

Conflict of interest statement

There is no conflict of interest.

Acknowledgments

This research is supported by the U.S. NSF grant DMS-1120824, China NSF grant 31271408, and Tsinghua university start up funding.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.gene.2013.07.092.

References

Almeida, J.S., Carrico, J.A., Maretzek, A., Noble, P.A., Fletcher, M., 2001. Analysis of genomic sequences by chaos game representation. Bioinformatics 17, 429–437.
Bateman, A., et al., 2002. The Pfam protein families database. Nucleic Acids Res. 30 (1), 276–280.
Berry, M.W., Drmac, Z., Jessup, E.R., 1999. Matrices, vector spaces, and information retrieval. SIAM Rev. 41 (2), 335–362.
Cao, Y., et al., 1998. Conflict among individual mitochondrial proteins in resolving the phylogeny of eutherian orders. J. Mol. Evol. 47 (3), 307–322.
Carr, K., Murray, E., Armah, E., He, R.L., Yau, S.S.-T., 2010. A rapid method for characterization of protein relatedness using feature vectors. PLoS One 5 (3), e9550.
Chan, R.H., Wang, R.W., Yeung, H.M., 2010. Composition vector method for phylogenetics — a review. Proc. 9th Int. Symp. Operations Research and Its Applications. ORSC & APORC, Chengdu, China, pp. 13–20.
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. Bioinform. 9 (1), 79–87.
Chou, K.C., 2011. Some remarks on protein attribute prediction and pseudo amino acid composition (50th anniversary year review). J. Theor. Biol. 273, 236–247.
Chou, K.C., Shen, H.B., 2009. Review: recent advances in developing web-servers for predicting protein attributes. Nat. Sci. 2, 63–92.
Chu, K.H., Qi, J., Yu, Z.G., Anh, V., 2004. Origin and phylogeny of chloroplasts: a simple correlation analysis of complete genomes. Mol. Biol. Evol. 21, 200–206.
Davies, M.N., Secker, A., Freitas, A.A., Timmis, J., Clark, E., Flower, D.R., 2008. Alignment-independent techniques for protein classification. Curr. Proteomics 5, 217–223.
Deng, M., Yu, C., Liang, Q., He, R.L., Yau, S.S.-T., 2011. A novel method of characterizing genetic sequences: genome space with biological distance and applications. PLoS One 6 (3), e17293.
Eckart, C., Young, G., 1936. The approximation of one matrix by another of lower rank. Psychometrika 1, 211–218.
Fauchere, J., Pliska, V., 1983. Hydrophobic parameters of amino acid side chains from the partitioning of N-acetyl-amino-acid amides. Eur. J. Med. Chem. 18, 369–375.
Gao, L., Qi, J., 2007. Whole genome molecular phylogeny of large dsDNA viruses using composition vector method. BMC Evol. Biol. 7, 41. http://dx.doi.org/10.1186/1471-2148-7-41.
Gotoh, O., 1982. An improved algorithm for matching biological sequences. J. Mol. Biol. 162, 705–708.
Huang, Y., Cai, J., Ji, L., Li, Y., 2004. Classifying G-protein coupled receptors with bagging classification tree. Comput. Biol. Chem. 28, 275–280.
Li, M., Badger, J.H., Chen, X., Kwong, S., Kearney, P., Zhang, H., 2001. An information based sequence distance and its application to whole mitochondrial genome phylogeny. Bioinformatics 17, 149–154.
Liao, B., Wang, T., 2004. New 2D graphical representation of DNA sequences. J. Comput. Chem. 25, 1364–1368.
Liu, L., Ho, Y., Yau, S.S.-T., 2006. Clustering DNA sequences by feature vectors. Mol. Phylogenet. Evol. 41, 64–69.
Lo Conte, L., Ailey, B., Hubbard, T.J., Brenner, S.E., Murzin, A.G., Chothia, C., 2000. SCOP: a structural classification of proteins database. Nucleic Acids Res. 28 (1), 257–259.
Lu, G., Zhang, S., Fang, X., 2008. An improved string composition method for sequence comparison. BMC Bioinforma. 9 (Suppl. 6), S15.
Minajeva, A., Kulke, M., Fernandez, J.M., Linke, W.A., 2001. Unfolding of titin domains explains the viscoelastic behavior of skeletal myofibrils. Biophys. J. 80 (3), 1442–1451.
Needleman, S.B., Wunsch, C.D., 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol. 48, 443–453.
Otu, H.H., Sayood, K., 2003. A new sequence distance measure for phylogenetic tree construction. Bioinformatics 19 (16), 2122–2130.
Pham, T.D., Zuegg, J., 2004. A probabilistic measure for alignment-free sequence comparison. Bioinformatics 20 (18), 3455–3461.
Qi, J., Wang, B., Hao, B., 2004. Whole proteome prokaryote phylogeny without sequence alignment: a K-string composition approach. J. Mol. Evol. 58, 1–11.
Randic, M., Vracko, M., Lers, N., Plavsic, D., 2003. Novel 2-D graphical representation of DNA sequences and their numerical characterization. Chem. Phys. Lett. 368, 1–6.
Saitou, N., Nei, M., 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4 (4), 406–425.
Smith, T.F., Waterman, M.S., 1981. Identification of common molecular subsequences. J. Mol. Biol. 147, 195–197.
Stuart, G.W., Moffett, K., Leader, J.J., 2002a. A comprehensive vertebrate phylogeny using vector representations of protein sequences from whole genomes. Mol. Biol. Evol. 19, 554–562.
Stuart, G.W., Moffett, K., Baker, S., 2002b. Integrated gene and species phylogenies from unaligned whole genome protein sequences. Bioinformatics 18 (1), 100–108.
Takahashi, M., Kryukov, K., Saitou, N., 2009. Estimation of bacterial species phylogeny through oligonucleotide frequency distances. Genomics 93 (6), 525–533.
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.
Vinga, S., Almeida, J., 2003. Alignment free sequence comparison—a review. Bioinformatics 19, 513–523.
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.
Yau, S.S.-T., Wang, J., Niknejad, A., Lu, C., Jin, N., Ho, Y., 2003. DNA sequence representation without degeneracy. Nucleic Acids Res. 31, 3078–3080.
Yau, S.S.-T., Yu, C., He, R., 2008. A protein map and its application. DNA Cell. Biol. 27, 241–250.
Yu, C., Liang, Q., Yin, C., He, R.L., Yau, S.S.-T., 2010. A novel construction of genome space with biological geometry. DNA Res. 17, 155–168.
Yu, C., Deng, M., Yau, S.S.-T., 2011a. DNA sequence comparison by a novel probabilistic method. Inf. Sci. 181, 1484–1492.
Yu, C., Cheng, S.-Y., He, R.L., Yau, S.S.-T., 2011b. Protein map: an alignment-free sequence comparison method based on various properties of amino acids. Gene 486 (1–2), 110–118.
Yu, C., Deng, M., Cheng, S.-Y., Yau, S.-C., He, R.L., Yau, S.S.-T., 2013. Protein space: a natural method for realizing the nature of protein universe. J. Theor. Biol. 318, 197–204.