Gene 486 (2011) 110–118

Contents lists available at ScienceDirect

Gene

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

Protein map: An alignment-free sequence comparison method based on various properties of amino acids

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

[a] The Institute of Mathematical Sciences, The Chinese University of Hong Kong, Shatin, Hong Kong
[b] Department of Mathematics, Hong Kong University of Science and Technology, Hong Kong
[c] Department of Biological Sciences, Chicago State University, IL, USA
[d] Department of Mathematical Sciences, Tsinghua University, Beijing, PR China

ARTICLE INFO

Article history:
Accepted 9 July 2011
Available online 19 July 2011

Received by A.J. van Wijnen

Keywords:
Protein map
Classification
Phylogeny
Amino acid substitution
Multiple alignment
Mitochondrial genes

ABSTRACT

In this paper, we propose a new protein map which incorporates with various properties of amino acids. As a powerful tool for protein classification, this new protein map both considers phylogenetic factors arising from amino acid mutations and provides computational efficiency for the huge amount of data. The ten amino acid physico-chemical properties (the chemical composition of the side chain, two polarity measures, hydropathy, isoelectric point, volume, aromaticity, aliphaticity, hydrogenation, and hydroxythiolation) are utilized according to their relative importance. Moreover, during the course of calculation of genetic distances between pairs of proteins, this approach does not require any alignment of sequences. Therefore, the proposed model is easier and quicker in handling protein sequences than multiple alignment methods, and gives protein classification greater evolutionary significance at the amino acid sequence level.

1. Introduction

The proteins encoded in a genome, and how these genes are expressed, determine the material basis of an organism's anatomy and physiology. Today proteins are very diverse, and classifying them into a taxonomic order (protein classification) based on natural descent (homology), akin to the taxonomy of species, would be of great interest. During the process of evolution, life forms of increasing complexity have arisen, and these increases have involved the emergence of new proteins (Chothia and Gough, 2009). The discovery of new protein sequences is accelerating, but many of these proteins show similarity to existing amino acid sequences. Problems arise when detecting the similarity of proteins, and explaining their phylogenetic relations as well as when handling the huge amount of data.

From an evolutionary point of view, the amino acid sequences of proteins are now widely used to infer the phylogenetic relationships of species (Hashimoto and Hasegawa, 1996). Models of amino acid substitution are developed and used in the multiple alignment of protein sequences (Yang, 2006). These substitution models include empirical models (Dayhoff et al., 1978; Jones et al., 1992) and mechanistic models (Yang et al., 1998). The former describe the relative rates of substitution between amino acids without explicitly considering factors that influence the evolutionary process. They are often constructed by analyzing large quantities of sequence data. The latter, on the other hand, consider the biological process involved in amino acid substitution, such as mutational biases in DNA, translation of the codons into amino acids, and acceptance or rejection of the resulting amino acid after filtering by natural selection. Some of the mechanistic models incorporate physico-chemical properties of amino acids (such as size and polarity) (Yang et al., 1998). Thus, amino acid replacement models determine the similarity of protein sequences by representing more phylogenetic information. This similarity has strong biological evolutionary significance. However, multiple alignment of the sequences is required to use these models. This turns out to be computationally difficult with large biological databases.

From a computational point of view, on the other hand, alignment-free approaches have been developed to overcome the 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,

Abbreviations: MSA, multiple sequence alignment; CV, composition vector; UPGMA, unweighted pair-group method with arithmetic means; EIIP, electron-ion interaction potential.

* Corresponding author at: Department of Mathematical Sciences, Tsinghua University, Beijing,100084, PR China. Tel./fax: +1 312 996 3065.
E-mail address: yau@uic.edu (S.S.-T. Yau).

0378-1119/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2011.07.002

2004; Yu et al., 2011). Furthermore, sequence vector representation approaches without alignment are also prevalent, such as composition vector (Chan et al., 2010; Chu et al., 2004; Gao and Qi, 2007; Qi et al., 2004), feature vector (Carr et al., 2010; Liu et al., 2006), and moment vector (Deng et al., 2011; Yu et al., 2010). However, all of these methods treat the DNA or protein sequences as simple character strings, and then consider occurrence frequencies and distribution of nucleotides or amino acids, or complexity of sequences and so on. Thus these alignment-free approaches, without considering biological factors that influence the evolutionary process, may have diluted the phylogenetic information.

In fact, the physico-chemical properties of amino acids are found to have strong effects on amino acid substitution rates (Xia and Li, 1998; Yang et al., 1998). Thus these properties directly determine the estimation of distance between two amino acid sequences. In our previous work (Yau et al., 2008), we constructed a protein map, which can be used to specify the similarity of two proteins without sequence alignment. More explicitly, we mathematically prove that the correspondence between moment vectors and protein sequences is one-to-one, and the distance between two amino acid sequences is defined by the Euclidean distance between two corresponding moment vectors. Although we consider the biological factors involved in amino acid substitution — the amino acid hydrophobicity scale values are used to construct the moment vector of protein sequence, other important physico-chemical properties of amino acids are not involved.

In this paper, we propose a new protein map which takes various properties of amino acids into account. The addition of amino acid properties to the protein map means that consideration is given to phylogenetic factors arising from amino acid mutations without sacrificing computational efficiency. The 10 amino acid properties (the chemical composition of the side chain, two polarity measures, hydropathy, isoelectric point, volume, aromaticity, aliphaticity, hydrogenation, and hydroxythiolation) are used according to their relative importance (Xia and Li, 1998). Moreover, during the course of calculation of genetic distances between pairs of proteins, we do not need to perform any sequence alignment. Therefore, the proposed model is easier and quicker in handling protein sequences than multiple alignment methods, and makes the protein classification have more evolutionary significance by including information from the amino acid sequence level.

2. Materials and methods

2.1. The protein map

In this section, we present a brief overview about our previous work — protein map (Yau et al., 2008). Firstly, we construct a protein sequence graphical representation on two quadrants of the Cartesian coordinate system. The vectors corresponding to the 20 amino acids lie on the line segment whose x-coordinate value is 1 and whose y-coordinate values are between −1 and 1. The y-coordinate values of the 20 amino acid vectors are all distinct (see Fig. 1). The ordering of these y-coordinate values is according to amino acid hydrophobicity scale values. Then we obtain the graphical representation of amino acid sequence based on this vector system. The points in the graphical representation are obtained by the sum of vectors representing amino acids in the sequence. For example, in Fig. 2, we give the graphical representation of the first 10 vectors of human beta-globin amino acid sequence on the vector system shown in Fig. 1, and the graphical representation of the whole human beta-globin amino acid sequence based on the same vector system is also shown in Fig. 3. Here we should point out that the protein sequence graphical representation has no circuits or degeneracy, and the correspondence between the sequence and the graphical curve can be mathematically proven to be one-to-one. This is why the y-coordinate values of the 20 amino acid vectors must be all distinct. Secondly, given the graphical curve of a protein sequence that can be represented by a sequence of points $(1,y_1)$, $(2,y_2)$, …, (n,y_n), we construct the moments as follows:

$$M_j = \sum_{i=1}^{n} \frac{(x_i - y_i)^j}{n^j}, \quad j = 1, 2, \ldots, n, \tag{1}$$

where n is the number of amino acids contained in a protein sequence, and (x_i, y_i) represents the position of the ith amino acid in the protein graphical curve. Thus each protein sequence has an n-dimensional moment vector $(M_1, M_2, \ldots, M_n)$. We again can give a rigorous proof to show that the correspondence between a protein sequence having n amino acids and its n-dimensional moment vector $(M_1, M_2, \ldots, M_n)$ is one-to-one (see theorem from Yau et al., 2008). Finally, by using these moment vectors, we construct a protein map (actually, it is a two-dimensional plane with the first two momentsM_1 and M_2 being x-axis and y-axis, respectively). Each protein sequence can be represented as a point in this map. Here we should emphasize that we only need to calculate the first two moments (M_1,M_2) to determine the biological information of protein sequences. Remember that in the central limit theorem in probability and statistics, the limiting process is Gaussian. For a Gaussian distribution, the first two moments determine the density function. Thus, theoretically, the first several moments play a critical role in the n-dimensional moment vectors. Moreover, the analysis using real protein data shows that the first two or three moments have already characterized the protein sequence sufficiently well. The genetic (evolutionary) distance between two proteins can be obtained through the Euclidean distance between the corresponding points in this protein map.

The only biological factor involved in the construction process of a protein map is the hydrophobicity property of amino acids. The other important biological elements involved in amino acid substitution are not considered. Therefore, we will develop in this paper a new protein map that takes into account more biological factors that shape protein evolution at the amino acid level.

2.2. The new protein map

The physico-chemical properties of amino acids are found to have strong effects on amino acid substitution rates and pattern of protein evolution (Xia and Li, 1998; Yang et al., 1998), so we need to consider more amino acid properties. In Xia and Li's (1998) work, 10 amino acid properties (chemical composition of the side chain, two polarity measures, hydropathy, isoelectric point, volume, aromaticity, aliphaticity, hydrogenation, and hydroxythiolation) are studied. Furthermore, the authors evaluate the relative importance of amino acid properties with respect to (1) the evolution of the genetic code, (2) the amino acid composition of proteins, and (3) the pattern of nonsynonymous substitutions. Their work provides us with a strong basis for constructing a new protein map.

In the second column of Table 1 in this paper and other 9 tables in the Supplementary Data, we give the values of 10 amino acid properties which Xia and Li (1998) studied: chemical composition of the side chain, polarity, volume (Grantham, 1974), polar requirement (Woese et al., 1966), hydropathy (Kyte and Doolittle, 1982), isoelectric point (Alff-Steinberger, 1969), PC I (aliphaticity), PC II (hydrogenation), PC III (aromaticity), and PC IV (hydroxyethylation). The last four properties are from Sneath's (1966) paper and each of them is a principal component, which represents major composite chemical factors.

In the third column of Table 1 in text and other 9 tables in Supplementary Data, we give the y-coordinate values of amino acids which are used to construct the new protein map. The y-coordinate value of amino acid with the largest property value is assigned to be 1, and with the smallest property value is assigned to be −1. The y-coordinate values of other amino acids are between −1 and 1.

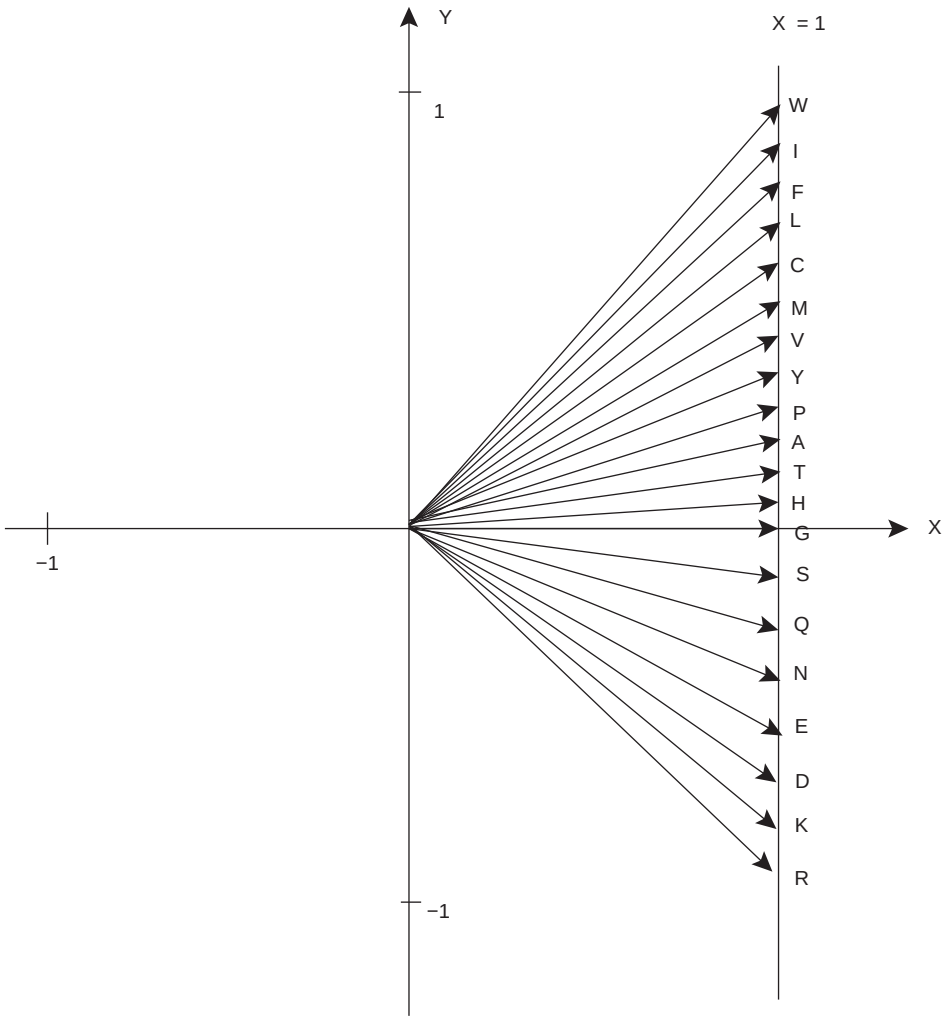

Fig. 1. An amino acid vector system based on the hydrophobicity scale values.

The *y*-coordinate difference between every two amino acids is proportional to the difference between their corresponding property values. For example, in Table 1, Arg (R) has the largest isoelectric point value (10.76) and Asp (D) has the smallest isoelectric point value (2.77). So, their *y*-coordinate values are assigned to be 1 and −1, respectively. For the 20 amino acids, the *y*-coordinate difference between every two of them is proportional to the difference between the corresponding property values of these two. For the calculation details, please see Supplementary Data. There are many amino acids which have the same values for some property. For example, in S. Table 1 in Supplementary Data, Leu (L), Ala (A), Val (V), Ile (I), Phe (F), and Met (M) have the same value zero with the property "chemical composition of the side chain". In this case, we add a very small perturbation difference value (0.001) for their *y*-coordinate values: Phe (F) with 0.001, Ile (I) with 0.002, Val (V) with 0.003, Ala (A) with 0.004, and Leu (L) with 0.005. The purpose for this is that we guarantee

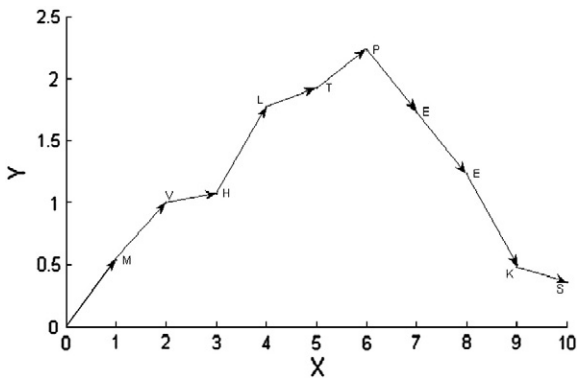

Fig. 2. Graphical representation of first 10 amino acids of human beta-globin sequence based on the vector system of Fig. 1.

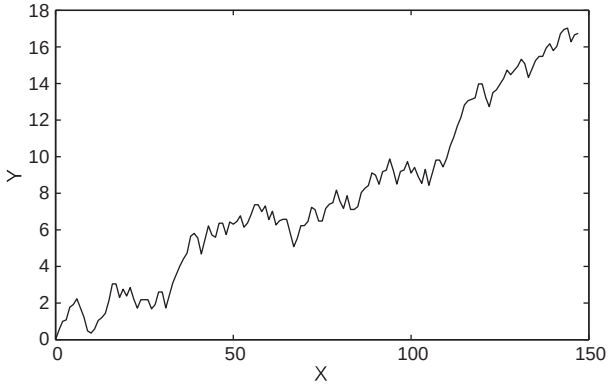

Fig. 3. Graphical representation of complete human beta-globin amino acid sequence based on the vector system of Fig. 1.

Table 1
Isoelectric point values (Alff-Steinberger, 1969) and *y*-coordinate values of 20 amino acids.

Amino acid	Isoelectric point	*y*-coordinate value
Arg (R)	10.76	1
Lys (K)	9.74	0.74468
His (H)	7.59	0.20651
Pro (P)	6.30	−0.1164
Thr (T)	6.16	−0.15144
Ile (I)	6.02	−0.18648
Ala (A)	6.00	−0.19149
Leu (L)	5.98	−0.1965
Gly (G)	5.97	−0.199
Val (V)	5.96	−0.2015
Trp (W)	5.89	−0.21902
Met (M)	5.74	−0.25657
Ser (S)	5.68	−0.27159
Tyr (Y)	5.66	−0.2766
Gln (Q)	5.65	−0.2791
Phe (F)	5.48	−0.32165
Asn (N)	5.41	−0.33917
Cys (C)	5.07	−0.42428
Glu (E)	3.22	−0.88736
Asp (D)	2.77	−1

the *y*-coordinate values of the 20 amino acid vectors are all distinct. This is crucial for the unique correspondence of moment vectors as we explained in the previous section.

For one amino acid sequence, we can obtain 10 graphical representations by using these 10 vectors (10 tables). For each graphical representation, we can get one moment vector according to our protein map method. Thus, for one protein sequence with n amino acids, we have 10 moment vectors associated with it:

$$\left(M_{1,1}, M_{1,2}, \ldots, M_{1,n}\right), \left(M_{2,1}, M_{2,2}, \ldots, M_{2,n}\right), \ldots, \left(M_{10,1}, M_{10,2}, \ldots, M_{10,n}\right).$$

Since the first several moments of one n-dimensional moment vector are crucial as we discussed before, we choose the first 3 moments for each vector (the subsequent demonstration for real mitochondrial gene data shows that this choice is appropriate). Therefore, for one amino acid sequence, we have a 30-dimensional combined vector associated with it:

$$\left(M_{1,1}, M_{1,2}, M_{1,3}, M_{2,1}, M_{2,2}, M_{2,3}, \ldots, M_{10,1}, M_{10,2}, M_{10,3}\right) \quad (2)$$

Up to now, we have fully utilized the values of 10 physico-chemical properties of amino acids. However, the relative importance of amino acid properties for determining the substitution rate and pattern of protein evolution may be quite different (Xia and Li, 1998). Therefore, it is necessary to add weights to the components of 30-dimensional vector according to their significance. In Xia and Li's (1998) article, after studying 10 protein-coding mitochondrial genes from 19 mammalian species, the authors pointed out that the genetic code appears to have evolved toward minimizing polarity and hydropathy but not the other seven properties; only the chemical composition and isoelectric point appear to have affected the amino acid composition of the proteins studied; all 10 properties except for PC IV affect the rate of amino acid nonsynonymous substitution. Furthermore, the authors give the effects of amino acid properties on the rates of amino acid substitution by numerical values. The last column (mean%) of Table 5 in their article (Xia and Li, 1998) is the average of percentage changes over the 10 genes for each of the 10 amino acid properties. The larger the percentage is, the more important the property is. Now we use these mean values (%) as the weights to improve the 30-dimensional vector. Although Xia and Li (1998) claimed that PC IV does not affect the rate of amino acid substitution, we still assign it with some small weight (to be 0.1). Thus, for one protein sequence, we have a weighted 30-dimensional vector associated with it, as follows:

$$\begin{aligned}&(0.2909M_{1,1}, 0.2909M_{1,2}, 0.2909M_{1,3}, 0.324M_{2,1}, 0.324M_{2,2}, 0.324M_{2,3},\\ &0.299M_{3,1}, 0.299M_{3,2}, 0.299M_{3,3}, 0.3749M_{4,1}, 0.3749M_{4,2}, 0.3749M_{4,3},\\ &0.2358M_{5,1}, 0.2358M_{5,2}, 0.2358M_{5,3}, 0.4348M_{6,1}, 0.4348M_{6,2}, 0.4348M_{6,3},\\ &0.2238M_{7,1}, 0.2238M_{7,2}, 0.2238M_{7,3}, 0.1736M_{8,1}, 0.1736M_{8,2}, 0.1736M_{8,3},\\ &0.2819M_{9,1}, 0.2819M_{9,2}, 0.2819M_{9,3}, 0.1M_{10,1}, 0.1M_{10,2}, 0.1M_{10,3}) \quad (3)\end{aligned}$$

By using this weighted 30-dimensional vector, a new protein map is generated. In fact, this new protein map is a 30-dimensional moduli space of protein sequences. In this map, each point corresponds to an amino acid sequence. The natural distance between two points in the map reflects the biological (evolutionary) distance between these two protein sequences. This allows us to perform the protein classification analysis at the amino acid sequence level. The evolutionary distance between two protein sequences obtained by this new protein map embodies the phylogenetic information with amino acid substitution rates, thus it has more biological evolutionary significance (as it is demonstrated later). Furthermore, in the next section, we will see that the computational time for this distance is considerably shorter than by alignment method.

3. Results

To test that the new protein map obtained in this way truly incorporates the classification and phylogenetic analysis, we apply it to the real protein sequence data. The first data set consists of the 10 proteins (COI, COIII, COII, Cyt-b, ND1, ATPase 6, ND4, ND5, ND6, ND2) encoded by the same strand of the mitochondrial genome from 19 mammalian species. In the 13 protein-coding mitochondrial genes, only 10 were used, with the 3 shortest genes (ATPase 8, ND3, and ND4L) excluded. The 10 proteins are concatenated into one long amino acid sequence and analyzed as one protein sequence. This data set is directly obtained from Xia and Li (1998) paper. The 19 mammalian species are hedgehog (GenBank accession number X88898), mouse (J01420), rat (X14848), cat (U20753), gray seal (X72004), harbor seal (X63726), horse (X79547), donkey (X97337), rhinoceros (X97336), cow (V00654), fin whale (X61145), blue whale (X72204), gibbon (X99256), Sumatran orangutan (X97707), Bornean orangutan (D38115), gorilla (X93347), pygmy chimpanzee (D38116), chimpanzee (D38113), and human (X93334).

Now we have 19 long concatenated amino acid sequences which are from the mitochondrial genomes of 19 mammalian species. For each amino acid sequence we calculate the weighted 30-dimensional moment vector (Eq. (3)) as shown in section "The new protein map". In this way, we obtain 19 weighted 30-dimensional moment vectors for these 19 mammalian species. By computing the Euclidean distance between pairs of these vectors, we get the distance matrix for species. Since these genetic sequences do not have large divergences, we use UPGMA method (Sokal and Sneath, 1963), which generates the rooted trees, to reconstruct the phylogenetic tree (Fig. 4). All the resulting trees in this paper are reconstructed by MEGA software (Kumar et al., 2008). The evolutionary relationships of primates (Gibbon, Gorilla, Human, Chimp, Pygmy Chimp, Sumatran Orang, and Bornean Orang) in our result are the same as those in Xia and Li's result (Fig. 1 from Xia and Li, 1998). The groups of perissodactyls (Rhino, Donkey, and Horse), carnivores (Cat, Gray Seal, and Harbor Seal), artiodactyls (Cow, Baleen Whale, and Fin Whale), and rodents (Rat and Mouse) are also clearly differentiated in our result. In fact, the sister taxa relationships of these groups are controversial (Cao et al., 1998; Janke et al., 1997; Xia and Li, 1998). One big difference between our result and Xia and Li's result is the position of hedgehog in the phylogenetic tree. Our result suggests that the insectivore hedgehog was the earliest-diverging of these

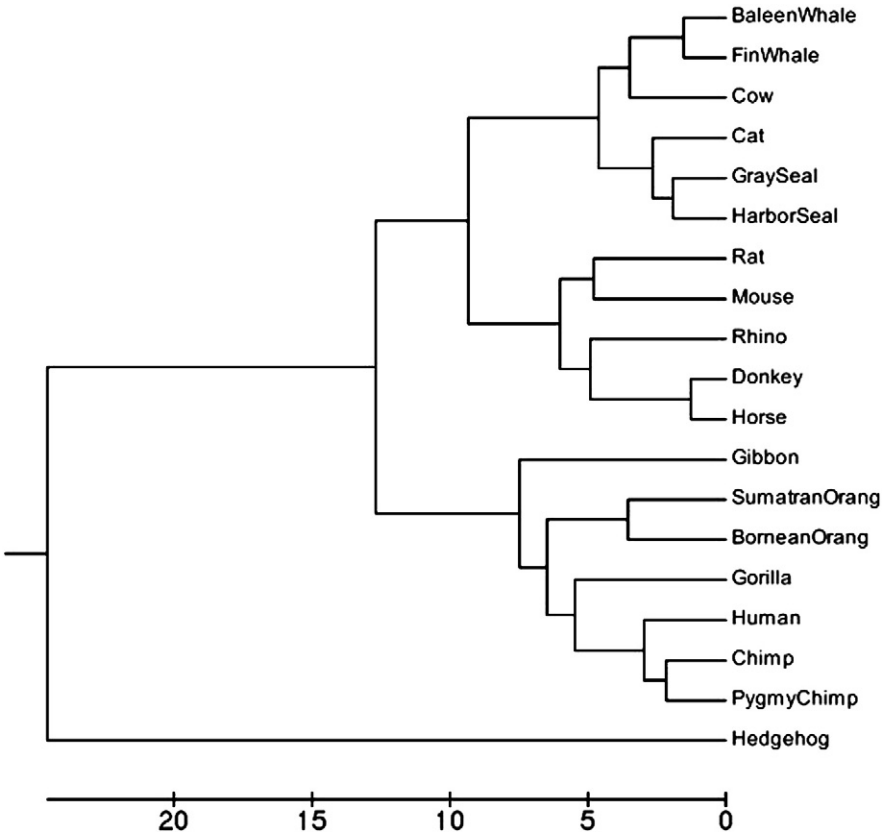

Fig. 4. The phylogenetic tree for the 19 mammalian species based on our new protein map.

eutherian sequences. This suggestion coincides with those found by Krettek et al. (1995) and Cao et al. (1998). The branch length in Xia and Li's unrooted phylogenetic tree (Xia and Li, 1998) is proportional to the number of nonsynonymous codon substitutions. The branch length in our result, on the other hand, is the Euclidean distance between two points corresponding to two amino acid sequences, in the new protein map.

3.1. Comparison with multiple sequence alignment (MSA) method

In order to show the computational efficiency of our approach we use the existing alignment tool ClustalW to make the multiple alignment for these 19 protein sequences. We choose the BLOSUM30 amino acid substitution matrix in this process. It took us about 4.5 min to get the alignment result on our Intel(R) Core(TM)2 Duo CPU E8400@ 3.00 GHz, 2.99 GHz Windows PC with 1.96 GB RAM. However, by our new approach, we obtain the distance matrix in only 20.224729s by a Matlab program on the same computer. The codes used to prepare this paper are available from the author upon request. Fig. 5 (A and B) shows two phylogenetic trees by the previous multiple alignment result with neighbor-joining and maximum parsimony methods. Both trees are reconstructed by MEGA software (Kumar et al., 2008) with bootstrap support analysis. The neighbor-joining tree (Fig. 5A) also suggests that the hedgehog was the earliest-diverging of these eutherian sequences, but the maximum parsimony tree (Fig. 5B) shows that hedgehog has the close relationship with rodents. As we discussed before, our result suggests hedgehog has the earliest-diverging position, although this point is controversial (Lawn et al., 1997). Furthermore, the maximum parsimony tree fails to cluster the carnivores group (Cat, Gray Seal, and Harbor Seal). Thus, our approach surpasses the multiple alignment method for both computational efficiency and biological results.

3.2. Comparison with another alignment-free method

Recently, Hao Bailin's laboratory developed a K-mer-based composition vector (CV) method with subtracted background "noise" modeled by a Markov chain estimator. By using this alignment-free approach, Hao's group obtained valuable results for both protein and genome sequences (Gao and Qi, 2007; Qi et al., 2004). However, for this method, when K becomes larger, the dimension of composition vector increases exponentially. For example, when $K=4$, the dimension of protein composition vector is $20^4=160{,}000$; when $K=5$, the dimension of protein composition vector is $20^5=3{,}200{,}000$. This brings great problems for computer memory. In Hao's method, the correlation $C(A, B)$ between two protein sequences A and B is calculated as the cosine function of the angle between the two corresponding vectors, and the distance $D(A, B)$ is defined as

$$D(A,B) = \frac{1-C(A,B)}{2} \tag{4}$$

We use a Matlab program of composition vector to deal with the 19 protein sequences. For $K=5$, it takes 126.296482s to obtain the distance matrix on the same PC computer. This takes much more time than our approach. Fig. 6 shows the neighbor-joining tree based on this distance matrix. We can see that the groups of primates, perissodactyls, carnivores, artiodactyls, and rodents are also clearly differentiated in this result. The position of hedgehog is closer to rodents. However, the position of Gibbon in primate group is different from the results of our method and MSA method. In this figure, the tree suggests orangutan was the earliest-diverging of these primate sequences. When we try to use $K=6$, the algorithm is out of memory for our PC computer. Actually, the dimension of protein composition vector for $K=6$ is $20^6=64{,}000{,}000$. This needs supercomputer with very large memory to handle. However,

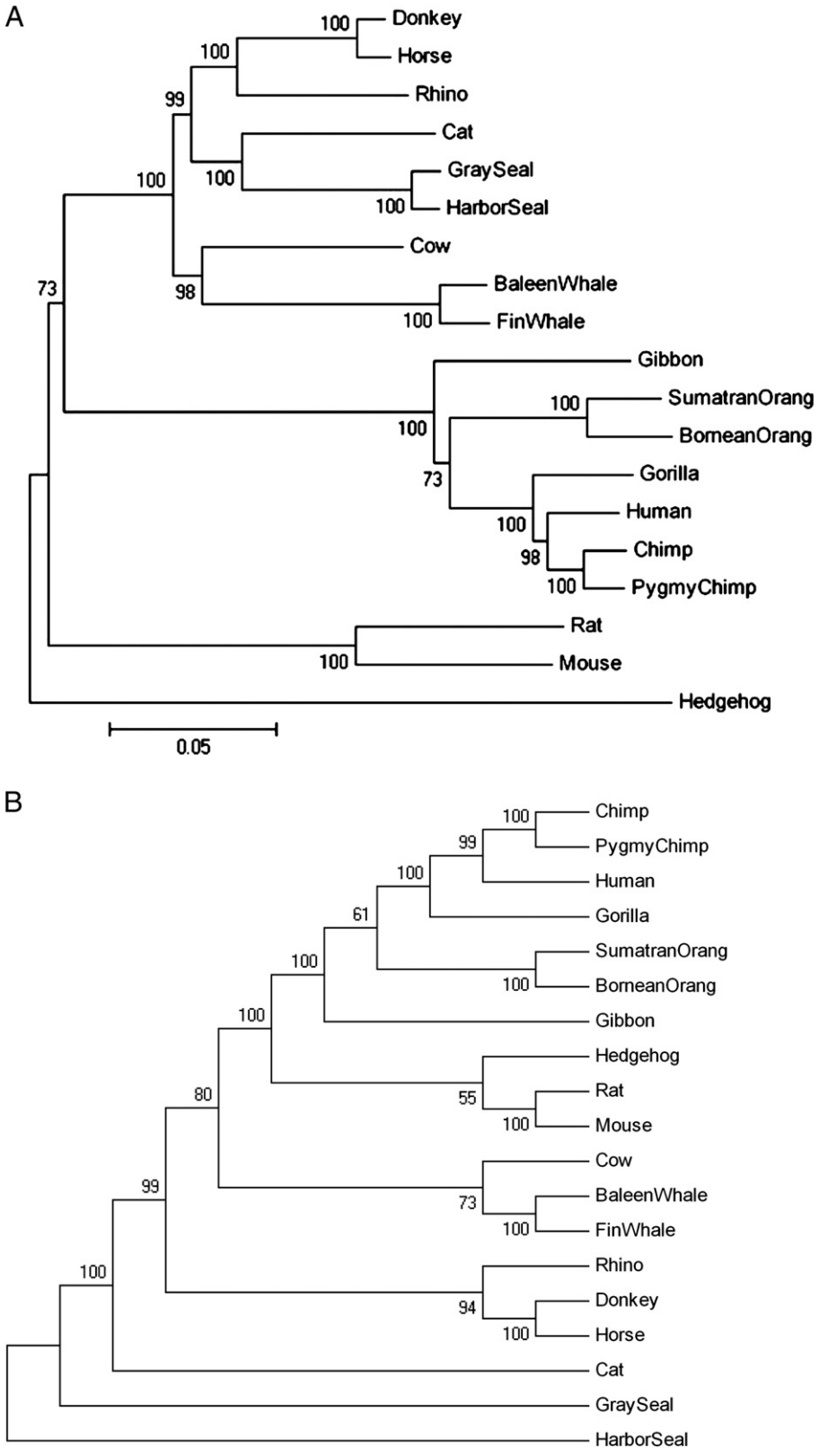

Fig. 5. A, the neighbor-joining tree for the 19 mammalian species based on multiple sequence alignment; B, the maximum parsimony tree for the 19 mammalian species based on multiple sequence alignment.

the weighted moment vector in our method is low dimensional (only 30 dimensions), and thus it does not require large computer memory to store. Even the ordinary PC computers can also run our algorithm. Therefore, our approach surpasses the composition vector method for both computational time and space.

In order to further illustrate the efficiency of our new protein map, we add some new mammalian mitochondrial proteins to the previous data set. They are tiger (EF551003), dog (U96639), wolf (EU442884), black bear (DQ402478), brown bear (AF303110), polar bear (AF303111), opossum (Z29573), wallaroo (Y10524), and platypus

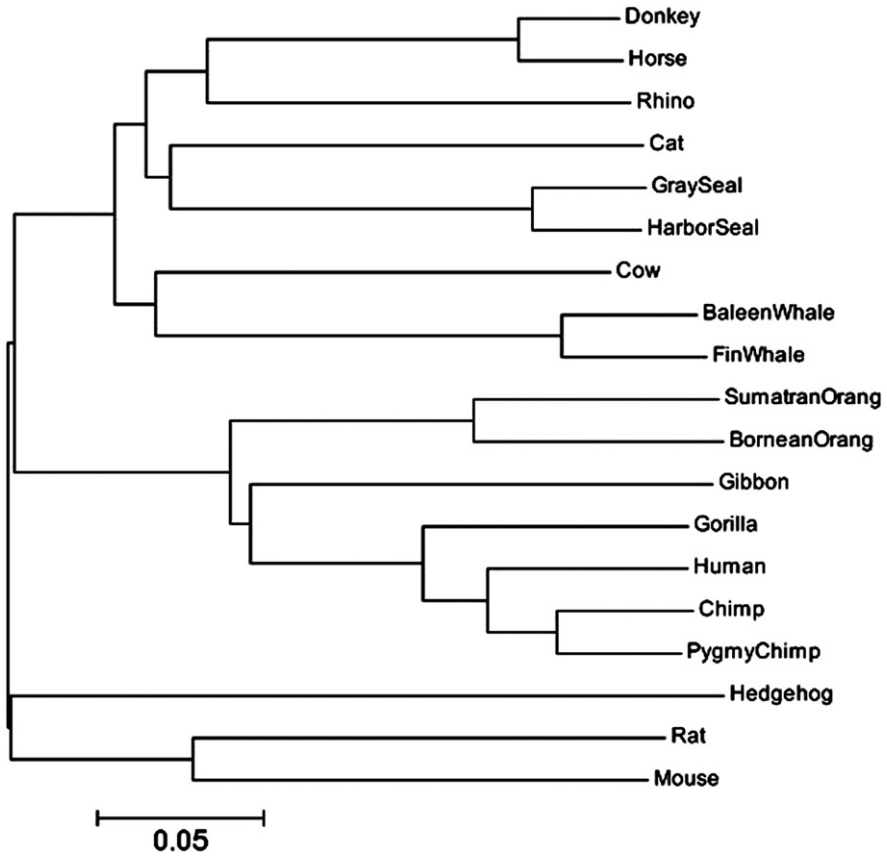

Fig. 6. The neighbor-joining tree for the 19 mammalian species based on Hao's composition vector method ($K=5$).

(X83427). Opossum, wallaroo, and platypus were mentioned but not used in Xia and Li's (1998) study because of difficulties in sequence alignment. As before, in the 13 protein-coding mitochondrial genes, only 10 were used, with the 3 shortest genes (ATPase 8, ND3, and ND4L) excluded. The 10 proteins are concatenated into one long amino acid sequence and analyzed as one protein sequence. For each amino acid sequence we calculate the weighted 30-dimensional moment vector as shown in section "The new protein map". Then we obtain 28 weighted 30-dimensional moment vectors for these 28 mammalian species. By computing the Euclidean distance between pairs of these vectors, we get the distance matrix for species and use UPGMA method (Sokal and Sneath, 1963) to reconstruct the phylogenetic tree (Fig. 7). In this figure, the Canis group (dog and wolf) and Ursidae group (black bear, brown bear, and polar bear) are clearly classified. Two marsupials (wallaroo and opossum) and one monotreme (platypus) are classified into one individual group because they are not eutherian mammalian species. The hedgehog is still far away from other clusters. Here we should point out that to get an accurate evolutionary tree for organisms, the complete genome sequences may be necessary. In this paper, we focus on the classification of protein sequences. Despite of this, this figure still clearly shows the similarity of these 28 protein sequences.

4. Discussion

The 10 amino acid properties and their relative importance we used in this work are taken directly from another paper (Xia and Li, 1998). But whether there are other important amino acid properties that affect protein evolution is uncertain. Recently, Pham (Pham, 2006) developed a LPC cepstral distortion measure for protein sequence comparison. In that method, the author also described some numerical values (EIIP) for the 20 amino acids, which make the protein sequence explicit for digital signal processing. However, the EIIP values for an amino acid are determined by a general pseudo-potential model, not by the direct physico-chemical experiments. Moreover, to our best knowledge, no literature evaluates whether these numerical values involve with protein evolution. In this paper, the relative importance of amino acid properties is evaluated based only on 10 protein-coding mitochondrial genes from 19 mammalian species, rather than a universal protein database. Thus, this relative importance may be limited only to mammalian mitochondrial proteins. Further investigation by considering more phylogenetic factors and a large protein database is desired. In addition, reported properties of amino acids may vary and depend on measurement procedures, which will introduce difficulties when selecting such data. For example, there are at least two different scales for hydrophobicity of the 20 amino acids (Fauchere and Pliska, 1983; Kyte and Doolittle, 1982). Therefore, these uncertain factors will affect our results. Further studies will be needed to decide suitable combination of amino acid properties with more biological information. In this work, the distance between two amino acid sequences is defined by the Euclidean distance between two corresponding moment vectors. Actually, we also use other distance measures such as Mahalanobis metric, Minkowski metric ($p>2$), cosine function of the angle between two vectors, etc. However, only the Euclidean distance gives us the convincing phylogenetic and clustering analysis results. Thus, only the Euclidean metric on the new protein map will give us the correct biological geometry.

5. Conclusion

In this paper, we have proposed a new protein map. This new protein map gives consideration to both phylogenetic factors with amino acid substitutions and computational efficiency for the huge number of protein sequences. The ten amino acid properties are well utilized according to their relative importance. During the course of calculation of genetic distances between pairs of proteins, we do not

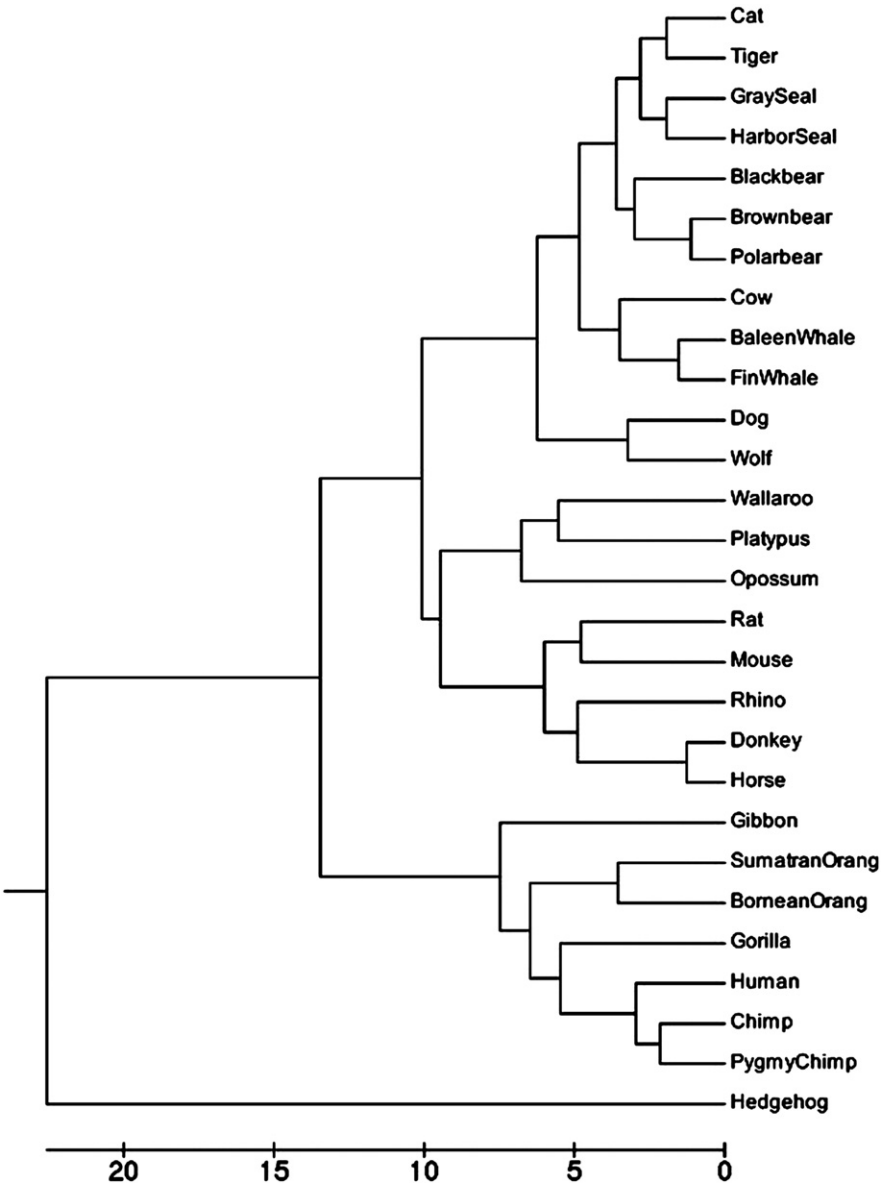

Fig. 7. The phylogenetic tree for the 28 mammalian species based on our new protein map.

need any alignment of sequences. Therefore, the proposed model is easier and quicker in handling protein sequences than multiple alignment methods, and gives protein classification greater evolutionary significance at the amino acid sequence level.

Acknowledgments

We thank the anonymous referees for providing us with constructive comments and suggestions. We also thank Dr. Max Benson for critically reading and editing the manuscript. The first author would like to thank Prof. Raymond Chan and Mr. Yeung Hau-Man at CUHK for sharing the Matlab code of Hao's composition vector method. This research was supported by NSF, Tsinghua University, and HKUST.

Appendix A. Supplementary data

Supplementary data to this article can be found online at doi:10.1016/j.gene.2011.07.002.

References

Alff-Steinberger, C., 1969. The genetic code and error transmission. Proc. Natl. Acad. Sci. U.S.A. 64, 584–591.
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.
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. doi:10.1371/journal.pone.0009550.
Chan, H.F., Wang, R.W., Wong, J.C., 2010. Maximum entropy method for composition vector method. In: Elloumi, M., Zomaya, A. (Eds.), Algorithms in Computational Molecular Biology: Techniques, Approaches and Applications. Wiley Series in Bioinformatics. doi:10.1002/9780470892107.ch27.
Chothia, C., Gough, J., 2009. Genomic and structural aspects of protein evolution. Biochem. J. 419 (1), 15–28.
Chu, K.H., Qi, J., Yu, Z.-G., Anh, V., 2004. Origin and phylogeny of chloroplasts revealed by a simple correlation analysis of complete genomes. Mol. Biol. Evol. 21 (1), 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.
Dayhoff, M.O., Schwartz, R.M., Orcutt, B.C., 1978. A model of evolutionary change in proteins. In: Dayhoff, M.O. (Ed.), Atlas of Protein Sequence and Structure, Vol 5. National Biomedical Research Foundation, Washington, D.C., pp. 345–352. Suppl. 3.
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. doi:10.1371/journal.pone.0017293.

Fauchere, J., Pliska, V., 1983. Hydrophobic parameters of amino-acid side-chains from the partitioning of N-acetylamino-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. doi:10.1186/1471-2148-7-41.
Grantham, R., 1974. Amino acid difference formula to help explain protein evolution. Science 185, 862–864.
Hashimoto, T., Hasegawa, M., 1996. Origin and early evolution of eukaryotes inferred from the amino acid sequences of translation elongation factors 1alpha/Tu and 2/G. Adv. Biophys. 32, 73–120.
Huang, Y., Cai, J., Ji, L., Li, Y., 2004. Classifying G-protein coupled receptors with bagging classification tree. Comput. Biol. Chem. 28, 275–280.
Janke, A., Xu, X., Arnason, U., 1997. The complete mitochondrial genome of the wallaroo (*Macropus robustus*) and the phylogenetic relationship among Monotremata, Marsupialia and Eutheria. Proc. Natl. Acad. Sci. U.S.A. 94, 1276–1281.
Jones, D.T., Taylor, W.R., Thornton, J.M., 1992. The rapid generation of mutation data matrices from protein sequences. CABIOS 8, 275–282.
Krettek, A., Gullberg, A., Arnason, U., 1995. Sequence analysis of the complete mitochondrial DNA molecule of the hedgehog, *Erinaceus europaeus*, and the phylogenetic position of the Lipotyphla. J. Mol. Evol. 41, 952–957.
Kumar, S., Nei, M., Dudley, J., Tamura, K., 2008. MEGA: a biologist-centric software for evolutionary analysis of DNA and protein sequences. Brief. Bioinform. 9, 299–306.
Kyte, J., Doolittle, R.F., 1982. A simple method for displaying the hydropathic character of a protein. J. Mol. Biol. 157, 105–132.
Lawn, R.M., Schwartz, K., Patthy, L., 1997. Convergent evolution of apolipoprotein(a) in primates and hedgehog. Proc. Natl. Acad. Sci. U.S.A. 94 (22), 11992–11997.
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.
Pham, T.D., 2006. LPC cepstral distortion measure for protein sequence comparison. IEEE Trans. Nanobioscience 5 (2), 83–88.
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.L., 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.
Sneath, P.H.A., 1966. Relations between chemical structure and biological activity. J. Theor. Biol. 12, 157–195.
Sokal, R.R., Sneath, P.H.A., 1963. Numerical Taxonomy. W. H. Freeman and Company, San Francisco, CA.
Vinga, S., Almeida, J., 2003. Alignment-free sequence comparison — a review. Bioinformatics 19 (4), 513–523.
Woese, C.R., Dugre, D.H., Dugre, S.A., Kondo, M., Saxinger, W.C., 1966. On the fundamental nature and evolution of the genetic code. Cold Spring Labor Symp. Quant. Biol. 31, 723–736.
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.
Xia, X., Li, W.H., 1998. What amino acid properties affect protein evolution? J. Mol. Evol. 47 (5), 557–564.
Yang, Z., 2006. Computational Molecular Evolution. Oxford University Press, New York.
Yang, Z., Nielsen, R., Hasegawa, M., 1998. Models of amino acid substitution and applications to mitochondrial protein evolution. Mol. Biol. Evol. 15 (12), 1600–1611.
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., 2011. DNA sequence comparison by a novel probabilistic method. Inf. Sci. 181, 1484–1492.