Comparison of genome sequences via projection extractor upon virtual mixer

Hongjie Yu Corresp., [1] , **Yuan-Ting Zhang** [1] , **Wei Fang** [1]

[1] Department of Mathematics, School of Information and Network Engineering, Anhui Science and Technology University, Bengbu, Anhui, People Republic of China

Corresponding Author: Hongjie Yu
Email address: yhj70@mail.ustc.edu.cn

To compare multiple genome sequences, we transform each primary genome sequence into corresponding *k*-mer-based vectors. According to the principle of independent component analysis (ICA), the operation can be regarded as mixing multiple source genomic signals via several sensors, through which we can obtain the mixed vectors with equal-length from the corresponding genome sequences with different length. However, this mixing operation is performed by counting all the *k*-mer-based frequencies, instead of using real hardware of sensors. Thus, we name this preprocessing operation as virtual mixer (VM). Using ICA-based transformation, we projected all the vectors upon their independent components to capture the coefficients-based feature vector through the projection extractor (PE), which has been proved to have a property of distance preserving. Then, we used the proposed VMPE model upon three representative real datasets of genome sequence to test the efficiency for the model. The contrastive analysis results indicate that the proposed VMPE model performs well in similarity analysis.

Comparison of genome sequences via projection extractor upon virtual mixer

Hong-Jie Yu, Yuan-Ting Zhang and Wei Fang
Department of Mathematics, School of Information and Network Engineering, Anhui Science and Technology University, China

Corresponding Author:
Hong-Jie Yu
1501 Huangshan Avenue, Bengbu, Anhui Province, 233000, P.R. China
Email address: yhj70@mail.ustc.edu.cn

ABSTRACT

To compare multiple genome sequences, we transform each primary genome sequence into corresponding *k*-mer-based vectors. According to the principle of independent component analysis (ICA), the operation can be regarded as mixing multiple source genomic signals via several sensors, through which we can obtain the mixed vectors with equal-length from the corresponding genome sequences with different length. However, this mixing operation is performed by counting all the *k*-mer-based frequencies, instead of using real hardware of sensors. Thus, we name this preprocessing operation as virtual mixer (VM). Using ICA-based transformation, we projected all the vectors upon their independent components to capture the coefficients-based feature vector through the projection extractor (PE), which has been proved to have a property of distance preserving. Then, we used the proposed VMPE model upon three representative real datasets of genome sequence to test the efficiency for the model. The contrastive analysis results indicate that the proposed VMPE model performs well in similarity analysis.

INTRODUCTION

In the field of bioinformatics, sequence comparison aims to discover relationship of similarity among various biological sequences. Sequence comparison of several genomes at the nucleotide sequence level can be accomplished by multiple sequence alignment (MSA). Recently, a web-based alignment services have been proposed (*Nguyen*, *2012*). Generally, based on an appropriate base-substitution model, the MSA are used to compute similarity scores. However, since species diverge widely over time, both genomic rearrangements and insertions/deletions make MSA a little difficult in genome comparison.

In fact, at present, the genomics necessarily requires approaches within the non-coding area of genomic sequence. Obviously, it can be used to compare the whole genomes (including coding and non-coding regions). It is an urgent need to develop a method which can be free of a specific gene set, and can also analyze nongenic regions. Later, alignment-free approaches are successively proposed to achieve this target (*Deng et al.*, *2011, Dong et al.*, *2018, Gao & Qi*, *2007, Vinga & Almeida*, *2003, Yu et al.*, *2010*). The representative method is involved in

frequency-based method called feature frequency profiles (FFP) (*Jun et al.*, *2010, Sims et al.*, *2009, Sims et al.*, *2009, Sims & Kim*, *2011*), which can be used as the comparison of whole genomes or genomic segments that may not be closely related and have latent remarkable rearrangement or have not shared a common set of genes, e.g. regulatory, intronic or nongenic regions.

As a classical alignment-free method, Blaisdell first introduced *k*-mer for biological sequence comparison (*Blaisdell*, *1986*). The relative approaches can be found in the reference (*Luczak et al.*, *2017*). Vinga reviewed more developments about alignment-free comparison (*Vinga & Almeida*, *2003*), which described many developed approaches to mining data and comparing multiple sequences. As a variation of approach for text comparison, Sims et al. (*Sims et al.*, *2009*) insist that, for every two different texts, the 'distance' between these two word frequency profiles can be regarded as a means of the dissimilarity between corresponding two texts. However, because there are no 'words' within a long string of base-pairs that constitute genome sequences, the differences among relative *k*-mer frequencies can be used to calculate distance values. For a given length sequence, the frequency information for all of the possible features (*k*-mers) is assembled into a FFP, where either the resolution or length of the features is the most important parameter.

The information theory (IT) has been widely used in the area of computational biology. Based on the use of digital signal processing (DSP) theory and algorithms, using genomic signal processing (GSP), one can analyze DNA or protein sequences. GSP methods convert DNA data to numerical values, thus it would offer the opportunity of applying existing DSP methods for genomic data. Examples of the use of GSP methods include performing cluster analysis of biological sequences and the K-means algorithm, proposing visualization method to inspect and analyze possible hidden behaviors (*Mendizabal-Ruiz et al.*, *2018*).

At the nucleotide, codon and amino acid levels, the researcher has derived optimal symbolic-to-digital mappings of the linear, nucleic acid strands into real or complex genomic signals (*Cristea*, *2002*). The proposed approach converts the sequences of polypeptides or nucleotides into digital genomic signals, and provides the possibility for using a large variety of signal processing approaches to their handling and analysis. It is also shown that using this representation one can better extracted some essential features of the nucleotide sequences.

Alignment-free sequence comparison and analysis greatly benefited from concepts which were derived from IT, such as mutual information and entropy. Within the review (*Vinga*, *2014*), the author investigated many aspects of IT applications, such as resolution-free metrics based on iterative maps, block-entropy estimation, prediction of transcription factor binding sites, comprising the classification of motifs and sequence characterization based on linguistic complexity and entropic profiles. As a function of the genomic location, the Entropic Profiler (EP) captures the essential region with respect to the whole genome (*Comin & Antonello*, *2013, Fernandes et al.*, *2009*).

A method (*Stuart et al.*, *2002*) have been developed to produce comprehensive gene and to reconstruct species phylogenies from the unaligned whole genome data, where one can use the

singular value decomposition (SVD) approach to character string frequencies analysis. Within the SVD-based dimension-reduced space, a quantitative comparison for the relative orientations of protein vectors provided straightforward and accurate estimations of sequence similarity, which can in turn be used to construct comprehensive gene trees. Later (*Stuart & Berry*, *2004*), using this SVD-based alignment-free method, the authors compared the predicted protein complement of 9 whole eukaryotic genomes ranging from yeast to man. They got simultaneous identification and definition of a large number of well conserved motifs and gene families. Meanwhile, a species tree was constructed, which supports one of two conflicting hypotheses for metazoan relationships. Based on SVD, the analysis of the entire protein complement of 9 whole eukaryotic genomes suggests that highly conserved motifs and gene families can be identified, and one can effectively compare within a single coherent definition space, where the extraction of gene and species trees can be easily implemented. The analysis can provide a basis for these definitions, when there is no explicit definition of orthologous or homologous sites.

Comon developed *Independent Component Analysis* (ICA) to find a linear representation of non-gaussian data so that all the components are statistically independent with each other, or as independent as possible (*Comon*, *1994*). Using such a representation, it seems to capture the essential structure of the data in several applications, including signal separation and feature extraction.

As an no-alignment method, the composition vector (CV) method (*Chan et al.*, *2012*) has been extensively studied recently. The abilities for ICA in feature extraction (*Huang & Zheng*, *2006, Yeredor*, *2002*), plus the successful use of composition vector or *k*-mer method in the comparison of sequences inspire us to combine them for improving the performance on similarity analysis.

In this study, we propose an ICA-based model for similarity analysis of genome sequences, i.e., virtual mixer & projection extractor model (VMPE), where we extracted the projections as the features for genome through optimizing the model. We also cluster the genome upon three real datasets, and visually inspect features of genome and analyze their similarities. Our results indicate the feasibility of applying the proposed method to compare genomes.

MATERIALS & METHODS

Model for the feature extraction from genome sequences

Through transforming the sequences of polypeptides or nucleotides into digital genomic signals, several methods offer the possibility to employ a large variety of signal processing approaches to deal with and analyze the sequences (*Cristea*, *2002*). In this section, we propose a novel model to extract information from genome sequences, and investigate the properties of the model.

In the proposed approach, the procedure is composed of four stages:

a) Transforming each primary genome sequence into corresponding k-mer-based vectors via our designed ‘virtual mixer’ (VM).

b) Projecting all these ‘mixed’ vectors upon independent component coordinate system to extract features through the hierarchy ‘projective extractor’ (PE);

c) Optimizing the number of segments s* for the best segmentation scheme; and d) Applying the final dimension-reduced feature vectors obtained by our hierarchy VMPE model on the similarities analysis among genome sequences. The application on real dataset demonstrated the validity of our proposed approach.

Preprocessing of sequences via K-mer-based virtual mixer (VM)

Hao Bailin's laboratory has proposed k-mer-based composition vector (CV) approach, where background 'noise' can be subtracted via a Markov chain estimator. Based on this no-alignment approach, the author obtained valuable results for both genome and protein sequences (*Gao & Qi, 2007, Qi et al., 2004*).

Description for k-mer approach

As for the k-mer approach, a description of the details can be described as follows. Let s be the primary genome sequence with length L, $\boldsymbol{s} = '\boldsymbol{N}_1\boldsymbol{N}_2\cdots\boldsymbol{N}_L'$, where $\boldsymbol{N}_l \in \{\mathrm{A},\mathrm{T},\mathrm{G},\mathrm{C}\}$, $l = 1,2,\cdots,L$.

Generally, a k-mer mode is a series of k consecutive characters within a sequence. The usual handling way to count k-mers within a sequence with length L is to use a sliding window with length k, shifting one frame base each time from position 1 to $L-k+1$, until all the entire genomes have been scanned. Thus, the k-mer-based feature vector can be denoted as:

$$\vec{\boldsymbol{m}}^{(i)} = \left(m_{i1}, m_{i2}, \cdots, m_{i,4^k}\right), \quad (1)$$

where m_{ij} denotes the frequency of the counting number for each corresponding pattern with k-consecutive characters, and i indicates the label of genome sequence. Meanwhile, all the $\vec{\boldsymbol{m}}^{(i)}$ have been normalized by its own length, i.e. $L-k+1$, which can keep the obtained feature vector be freed from the negative influence of different lengths for each genome sequence.

The number of the order for k-mer mode is just denoted as k. Since the parameter k has a great influence on the results of sequence analysis, it is a key to pick out an appropriate k. Some works have investigated the selection of k. For example, Wu et al. (*Wu et al., 2005*) proposed an optimal word size to calculate dissimilarity, where the optimal word size can be determined by length of sequence considered. Another solution was investigated by Sims et al. (*Sims et al., 2009, Sims et al., 2009*), who gave the lower limit and upper limit for the range of optimal length. The lower limit can be approximately determined by the average length of multiple sequences, named as $\overline{L}$. Then one can let k be the integer part of $log_4\left(\overline{L}\right)$.

Observing the mixed genomic signals from sequence

As proposed by Pierre Comon(*Comon, 1994*), Independent Component Analysis (ICA) can lessen the effects of noise or artifacts within the data as it focuses on separating a mixture of signals into their each different sources, respectively. ICA models observations as linear combinations of several certain components, or variables, which are selected as statistically independent as possible, i.e. the different components represent different non-overlapping information.

Thus, we can regard the feature vector $\vec{\boldsymbol{m}}^{(i)}$ in Eq. (1) as the mixture from several independent components (ICs). However, in this study, the course of mixture need not perform through real

hardware, i.e. just via virtual k-mer-based counter.

Assuming further that there are several virtual 'sensors' or virtual 'receivers', these sensors have different 'inherent frequencies' so that each 'sensor' records a mixture of the original source genomic signals with slightly different patterns.

The observed data m_{ij} are obtained through "sensor", i.e., virtual mixer (VM) that gives the mixing weights among several latent independent components. In the left side hand of Eq. (1), each observed vector $\vec{\boldsymbol{m}}^{(i)}$ can be expressed as a linear combination of n latent components $\vec{\boldsymbol{c}}^{(1)}$, $\vec{\boldsymbol{c}}^{(2)}$, …, and $\vec{\boldsymbol{c}}^{(n)}$, where i just denotes the label for the genome sequence. All that we directly observed are the n genomic signals, $\vec{\boldsymbol{m}}^{(1)}$, $\vec{\boldsymbol{m}}^{(2)}$, … , and $\vec{\boldsymbol{m}}^{(n)}$. So the combination can be represented as follows:

$$\vec{\boldsymbol{m}}^{(i)} = f_{i1} \cdot \vec{\boldsymbol{c}}^{(1)} + f_{i2} \cdot \vec{\boldsymbol{c}}^{(2)} + \cdots f_{in} \cdot \vec{\boldsymbol{c}}^{(n)}, \tag{2}$$

where f_{ij} are some coefficients that define the representation, $i = 1,2,\cdots,n$, and i is just the label for the original sequence or species. Also, $\vec{\boldsymbol{c}}^{(1)}$, $\vec{\boldsymbol{c}}^{(2)}$, …, and $\vec{\boldsymbol{c}}^{(n)}$ are as independent as possible with each other.

How can we estimate the coefficients f_{ij} in Eq. (2)? Collecting all these coefficients f_{ij} into a matrix $\boldsymbol{F}$, we want to use some general statistical properties to find the matrix $\boldsymbol{F}$ through which we can represent the observed genomic signals by several latent independent signals. Then, the question can just boil down to the very ICA-based problem which is started with how to find a good representation of multivariate data.

Extracting features from mixed genomic signals

However, a general solution to the problem can be found by using independent component analysis (ICA) upon the observed data $\vec{\boldsymbol{m}}^{(1)}$, $\vec{\boldsymbol{m}}^{(2)}$, …, $\vec{\boldsymbol{m}}^{(n)}$, which are likewise collected into the k-mer-based matrix $\boldsymbol{K}$.

The mixed k-mer-based matrix $\boldsymbol{K}$ comprises n counting vectors as follows:

$$\begin{pmatrix} \vec{\boldsymbol{m}}^{(1)} \\ \vec{\boldsymbol{m}}^{(2)} \\ \vdots \\ \vec{\boldsymbol{m}}^{(n)} \end{pmatrix} = \begin{pmatrix} m_{11} & m_{12} & \cdots & m_{1,4^k} \\ m_{21} & m_{22} & \vdots & m_{2,4^k} \\ \vdots & \vdots & \ddots & \vdots \\ m_{n1} & m_{n2} & \cdots & m_{n,4^k} \end{pmatrix} \overset{\Delta}{=} \boldsymbol{K}, \tag{3}$$

where $i = 1,2,\cdots,n$, and $\vec{m}^{(i)}$ is a 4^k dimensional vector comprising of all the mixed elements transformed from the corresponding sequence $\boldsymbol{s}^{(i)}$ via the virtual mixer (VM)

Thus, we can estimate the coefficient matrix $\boldsymbol{F}$ for the genomic signals from original genome sequences (The elements f_{ij} in coefficient matrix $\boldsymbol{F}$ can be estimated by the developed algorithm, e.g. FastICA). Thus, Eq. (1) can be regarded as a statistical "latent variables" model. These latent components are assumed *unknown*, because we cannot know the values of c_{ij} without knowing all the properties of the virtual extractor system, which can be difficult in general. The projection coefficients f_{ij} are *unknown as well*, which we cannot record directly. The problem can then be

rephrased as that of how to determine the coefficients f_{ij} within matrix $\boldsymbol{F}$.

Moreover, it is usually more convenient to introduce a vector-matrix notation instead of all the n equations in Eq. (2). Let us denote the three series of variables by the corresponding row vectors, respectively. Thus, according to the block-matrix appearance, the relationship of $\boldsymbol{K}$ and $\boldsymbol{F}$ can be written as:

$$\begin{pmatrix} \vec{m}^{(1)} \\ \vec{m}^{(2)} \\ \vdots \\ \vec{m}^{(n)} \end{pmatrix} = \begin{pmatrix} \vec{f}^{(1)} \\ \vec{f}^{(2)} \\ \vdots \\ \vec{f}^{(n)} \end{pmatrix} \times \begin{pmatrix} \vec{c}^{(1)} \\ \vec{c}^{(2)} \\ \vdots \\ \vec{c}^{(n)} \end{pmatrix}, \tag{4}$$

where i is just the label for different genome sequence, $i = 1,2,\cdots,n$, and $\vec{\boldsymbol{c}}^{(i)}$ are n independent 4^k dimensional vectors which are obtained by using ICA upon all the n observed 4^k dimensional row vectors in matrix $\boldsymbol{K}$.

Virtual mixer & projection extractor (VM-PE)

VM-PE model for capturing genomic signals

Below are the descriptions for the VM-PE model, which is based on k-mer and ICA. Considering two transformations:

(a) $\boldsymbol{S} \xrightarrow{\tau_1} \boldsymbol{K}$

The matrix $\boldsymbol{S}$ denotes all the n involved primary genome sequences with different length, while the matrix $\boldsymbol{K} \in \mathcal{R}^{n \times 4^k}$ stands for the corresponding matrices which were mapped from all these n primary sequences. (Please see the first box shown in Fig. 1).

(b) $\boldsymbol{K} \xrightarrow{\tau_2} \boldsymbol{F}$

The matrix $\boldsymbol{F}$ comprises of all the n extracted feature vectors:

$$\vec{\boldsymbol{f}}^{(i)} = \left(f_{i1}, f_{i2}, \cdots, f_{in}\right), \tag{5}$$

which are transformed from k-mer-based mixer matrix $\boldsymbol{K}$ via ICA-based projection extractor (PE) (Please see the second box shown in Fig. 1).

Likewise, if we collect all the feature vectors into a matrix $\boldsymbol{F}$, then Eq. (5) can be rewritten as:

$$\begin{pmatrix} \vec{\boldsymbol{f}}^{(1)} \\ \vec{\boldsymbol{f}}^{(2)} \\ \vdots \\ \vec{\boldsymbol{f}}^{(n)} \end{pmatrix} = \begin{pmatrix} f_{11} & f_{12} & \cdots & f_{1n} \\ f_{21} & f_{22} & \vdots & f_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ f_{n1} & f_{n2} & \cdots & f_{nn} \end{pmatrix} \overset{\Delta}{=} \boldsymbol{F}, \tag{6}$$

Thus, the relationship between the matrices $\boldsymbol{K}$ and $\boldsymbol{F}$ shown in Fig. 1 can be described as:

$$\boldsymbol{K} = \boldsymbol{F} * \boldsymbol{C}. \tag{7}$$

That is,

$$\begin{pmatrix} m_{11} & m_{12} & \cdots & m_{1,4^k} \\ m_{21} & m_{22} & \cdots & m_{2,4^k} \\ \vdots & \vdots & \ddots & \vdots \\ m_{n1} & m_{n2} & \cdots & m_{n,4^k} \end{pmatrix} = \begin{pmatrix} f_{11} & f_{12} & \cdots & f_{1n} \\ f_{21} & f_{22} & \cdots & f_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ f_{n1} & f_{n2} & \cdots & f_{nn} \end{pmatrix} \times \begin{pmatrix} c_{11} & c_{12} & \cdots & c_{1,4^k} \\ c_{21} & c_{22} & \cdots & c_{2,4^k} \\ \vdots & \vdots & \ddots & \vdots \\ c_{n1} & c_{n2} & \cdots & c_{n,4^k} \end{pmatrix} \tag{8}$$

Obviously, the matrix $\boldsymbol{F}$ belongs to the space $\mathcal{R}^{n\times n}$, where the dimension of all the row vectors is uniform, i.e. n. Moreover, compared with matrix $\boldsymbol{K}$, it can be seen from Eqs. (7) or (8) that the dimension for feature matrix $\boldsymbol{F}$ is greatly reduced. Thus, the final feature matrix $\boldsymbol{F}$ deserves our special attention.

As shown in Fig. 1, we design a k-mer-based virtual mixer (VM) and ICA-based projection extractor (PE), through which we can extract the dimension-reduced essential features from the genome sequences.

Ensemble transformation

In fact, all the n obtained feature vectors $\vec{\boldsymbol{f}}^{(i)}$, $i=1,2,\cdots,n$, can be regarded as the projections, which are captured by projecting the k-mer-based 'observed variables' $\vec{\boldsymbol{m}}^{(i)}$ upon the n latent independent 4^k dimensional vectors $\vec{\boldsymbol{c}}^{(j)}$ in matrix $\boldsymbol{C}$, where f_{ij}, $i,j=1,2,\cdots,n$, are some real coefficients. Meanwhile, as the description of Eq. (2), all $\vec{\boldsymbol{c}}^{(i)}$ are statistically mutually independent.

In summary, as a whole, the compound transformation can be described as follows:

$$\mathcal{T}_2 \circ \mathcal{T}_1 : \boldsymbol{s}^{(i)} \mapsto \vec{\boldsymbol{f}}^{(i)} = \left(f_{i1}, f_{i2}, \cdots, f_{in}\right), \tag{9}$$

through which one can freely extract the features of the multiple genome sequences. Also, the ensemble transformation can be depicted as:

$$\textbf{\textit{Ker}}\ \varphi : \mathcal{S}^{n\times 1} \xrightarrow{\mathcal{T}} \mathcal{R}^{1\times\frac{n\cdot(n-1)}{2}} \tag{10}$$

where $\mathcal{S}^{n\times 1}$ stands for the *original string sequence space* comprising of n genome sequences with different length, while $\mathcal{R}^{1\times\frac{n\cdot(n-1)}{2}}$ denotes the *objective distance space* transformed from the *original space* via the proposed VM-PE scheme shown in Fig. 1.

Distance-preserving properties of transformation

As can be seen from the following proposition, the above-mentioned compound transformation embodies the essential property of the genome sequences. Thus, the φ can be interpreted as a kernel operator. Some properties for the proposed VM-PE algorithm can be depicted as:

***Definition* 1:** Within the *original string sequence space*, the distance between every two different sequences D($\boldsymbol{s}^{(i)}$, $\boldsymbol{s}^{(j)}$), is defined as:

$$\mathrm{D}\left(\boldsymbol{s}^{(i)}, \boldsymbol{s}^{(j)}\right) \overset{\text{def}}{=} corr\left(\vec{\boldsymbol{m}}^{(i)}, \vec{\boldsymbol{m}}^{(j)}\right) \tag{11}$$

where $\vec{\boldsymbol{m}}^{(i)}$ is the k-mer-based mixed vector for sequence $s^{(i)}$, and so is $\vec{\boldsymbol{m}}^{(j)}, i,j=1,2,\cdots,n$. Here,

the function $corr(\cdot,\cdot)$ denotes the correlation degree between two vectors, i.e., $\vec{\boldsymbol{m}}^{(i)}$ and $\vec{\boldsymbol{m}}^{(j)}$. Usually, it can be defined by the values of correlation coefficients from a pair of vectors.

***Definition* 2:** Let $\mathcal{R}^d$ be a *real normed space* with dimensions d, and let $\varphi: \mathcal{S}^{n\times 1} \mapsto \mathcal{R}^{1\times d}$ be a function from $\mathcal{S}^{n\times 1}$ to $\mathcal{R}^{1\times d}$, where $d = \frac{n\cdot(n-1)}{2}$. For any element within the space $\mathcal{S}^{n\times 1}$, such as $\boldsymbol{s}^{(i)}$ and $\boldsymbol{s}^{(j)}$, if $\mathrm{D}(\boldsymbol{s}^{(i)},\boldsymbol{s}^{(j)}) = \delta$ implies $\mathrm{D}(\varphi(\boldsymbol{s}^{(i)}),\varphi(\boldsymbol{s}^{(j)})) = \delta$, the function φ can be called δ-*distance preserving*.

***Theorem* 1:** $\mathcal{T}_2 \circ \mathcal{T}_1 : \boldsymbol{s}^{(i)} \mapsto \vec{\boldsymbol{f}}^{(i)} = (f_{i1}, f_{i2}, \cdots, f_{in})$ *is a distance-preserving transformation.*

***Proof*:** Since $\vec{\boldsymbol{m}}^{(i)}$ and $\vec{\boldsymbol{m}}^{(j)}$ are the *k*-mer-based mixed vectors of two corresponding sequences $\boldsymbol{s}^{(i)}$ and $\boldsymbol{s}^{(j)}$, respectively, $i, j = 1,2,\cdots,n$.

Let $\mathcal{T}_2 \circ \mathcal{T}_1$ be the compound transformation from the *original string sequence space* to *objective distance space*.

Then, the following equations:

$$\varphi(\boldsymbol{s}^{(i)}) = (\mathcal{T}_2 \circ \mathcal{T}_1)(\boldsymbol{s}^{(i)}) = \mathcal{T}_2[\mathcal{T}_1(\boldsymbol{s}^{(i)})] = \mathcal{T}_2(\vec{\boldsymbol{m}}^{(i)}) = \vec{\boldsymbol{f}}^{(i)} \tag{12}$$

will hold.

According to Eq. (2), we can rewrite them in summation form as follows:

$$\vec{\boldsymbol{m}}^{(i)} = \sum_{p=1}^{n} f_{ip} \cdot \vec{\boldsymbol{c}}^{(p)}, \tag{13}$$

while

$$\vec{\boldsymbol{m}}^{(j)} = \sum_{q=1}^{n} f_{jq} \cdot \vec{\boldsymbol{c}}^{(q)}. \tag{14}$$

Thus, there have

$$\mathrm{D}(\boldsymbol{s}^{(i)},\boldsymbol{s}^{(j)}) = \mathrm{D}(\vec{\boldsymbol{m}}^{(i)},\vec{\boldsymbol{m}}^{(j)}) = corr(\vec{\boldsymbol{m}}^{(i)},\vec{\boldsymbol{m}}^{(j)}) = corr\left(\sum_{p=1}^{n} f_{ip} \cdot \vec{\boldsymbol{c}}^{(p)}, \sum_{q=1}^{n} f_{jq} \cdot \vec{\boldsymbol{c}}^{(q)}\right)$$

$$= \sum_{p=1}^{n}\sum_{q=1}^{n}\left(corr\left(f_{ip} \cdot \vec{\boldsymbol{c}}^{(p)}, f_{jq} \cdot \vec{\boldsymbol{c}}^{(q)}\right)\right)$$

$$= \sum_{\substack{p=1 \\ p=q}}^{n}\left(corr\left(f_{ip} \cdot \vec{\boldsymbol{c}}^{(p)}, f_{jq} \cdot \vec{\boldsymbol{c}}^{(q)}\right)\right) + \sum_{\substack{p=1 \\ p\neq q}}^{n}\sum_{q=1}^{n}\left(corr\left(f_{ip} \cdot \vec{\boldsymbol{c}}^{(p)}, f_{jq} \cdot \vec{\boldsymbol{c}}^{(q)}\right)\right) \tag{15}$$

Since, every two independent component $\vec{\boldsymbol{c}}^{(p)}$ and $\vec{\boldsymbol{c}}^{(q)}$, $p \neq q$, are statistically mutually independent, then, according to the calculation for the correlation coefficients, it can be obtained that the second part within the right-hand side of Eq. (15) is just zero.

And hence, the whole Eq. (15) will be simplified into:

$$\mathrm{D}(\boldsymbol{s}^{(i)},\boldsymbol{s}^{(j)}) = \sum_{p=1}^{n} corr\left(f_{ip}\vec{\boldsymbol{c}}^{(p)}, f_{jp}\vec{\boldsymbol{c}}^{(p)}\right) = corr(\vec{\boldsymbol{f}}^{(i)},\vec{\boldsymbol{f}}^{(j)}), \tag{16}$$

where $i, j = 1,2,\cdots,n$.

By Eq. (12) and *Definition* 1, we have:

$$\mathrm{D}\left(\varphi(\boldsymbol{s}^{(i)}),\varphi(\boldsymbol{s}^{(j)})\right)=\mathrm{D}\left(\vec{\boldsymbol{f}}^{(i)},\vec{\boldsymbol{f}}^{(j)}\right)=corr\left(\vec{\boldsymbol{f}}^{(i)},\vec{\boldsymbol{f}}^{(j)}\right). \quad (17)$$

Comparing Eq. (16) with Eq. (17), by *Definition* 2, the following equation holds:

$$\mathrm{D}\left(\boldsymbol{s}^{(i)},\boldsymbol{s}^{(j)}\right)=\mathrm{D}\left(\varphi(\boldsymbol{s}^{(i)}),\varphi(\boldsymbol{s}^{(j)})\right),\ i,j=1,2,\cdots,n$$

Therefore, for two given genome sequences, $\boldsymbol{s}^{(i)}$ and $\boldsymbol{s}^{(j)}$, it can be seen that the above-mentioned ensemble transformation τ indeed has a property of δ-distance preserving. Here, superscript *i* or *j* denotes the label of sequences, and the sequences are usually non-equal length.

QED

According to *Theorem* 1 and *Definition* 2, we can calculate all the row vectors for matrix $\boldsymbol{F}$, such as $\vec{\boldsymbol{f}}^{(i)}=\left(f_{i1},f_{i2},\cdots,f_{in}\right)$, *i*=1, 2, …, *n*, where *i* denotes the label of *n* primary genome or proteome sequences. Then we can get *n* corresponding combinational coefficients vectors with the dimension *n*, which can be regarded as the features extracted from the original genome sequence via the proposed VM-PE algorithm, whose steps can be summarized as follows.

Input: multiple genome sequences with different length: $\boldsymbol{s}^{(1)}, \boldsymbol{s}^{(2)}, \ldots, \boldsymbol{s}^{(n)}$
begin
  for *i*=1 **to** *n* **do**
    Through virtual mixer (VE), transform each genome sequence $\boldsymbol{s}^{(i)}$ into a *k*-mer-based 4^k dimensional vector $\vec{m}^{(i)}$, which comprises *n* by 4^k observed matrix $\boldsymbol{K}$
  end for
  Using FastICA-based projection extractor (PE), factorize matrix $\boldsymbol{K}$ into independent component matrix $\boldsymbol{C}$ which is left multiplied by projection feature matrix $\boldsymbol{F}$
  for *i* =1 **to** *n*—1 **do**
    for *j*= *i*+1 **to** *n* **do**
      Calculate pairwise distances using $\boldsymbol{F}$ by $\mathrm{D}\left(\boldsymbol{s}^{(i)},\boldsymbol{s}^{(j)}\right)=\left\|\vec{\boldsymbol{f}}^{(i)}-\vec{\boldsymbol{f}}^{(j)}\right\|_2$
    end for
  end for
  for *i*=1 **to** *n* **do**
    The feature vector for the *i*-th genome sequence ← The *i*-th row vector of feature matrix $\boldsymbol{F}$
  end for
  Draw the dendrogram using the pairwise distances matrix
end

RESULTS

APPLICATION FOR THE REAL GENOME DATASET

We apply the proposed VMPE model upon the real genome dataset.

Data preparation

Table 1 shows the concise information for these 20 species in the GenBank. The first dataset includes the mitochondrial genome sequences from 20 eutherian species, which has also been investigated by several works (*Dai et al.*, *2011, Deng et al.*, *2011, Huang & Wang*, *2011, Huang et al.*, *2011, Yang et al.*, *2012, Yu & Huang*, *2012*).

Determining k^* for virtual mixer (VM)

Given multiple genome sequences with average length about 17,000 bps, the k-mer count vector $\vec{\boldsymbol{m}}^{(i)}$ becomes too sparse for $k \geq 8$ (*Yu*, *2013*). Thus, based on the higher order k-mer, the comparisons among multiple long sequences may not capture the essential feature of sequences. Therefore, we need to consider how to determine a rational order k^* for k-mers according to the approximate formula investigated in Ref. (*Sims et al.*, *2009*). Therefore, for the above-mentioned dataset, the integer k^* can be calculated as:

$$k^* \approx log_4(17000), \tag{18}$$

i.e. $k^*=7$.

Therefore, for the given genome sequence $\boldsymbol{s}^{(i)}$, $k^*=7$ is just the optimal order of k-mer for VMPE model at the stage of virtually mixing, where all the original sequences can be transformed into the corresponding mixed 4^7-dimensional vectors. In other words, through virtual mixer, all the obtained mixed 7-mer vector $\vec{\boldsymbol{m}}^{(i)}$ should uniformly be located at in dimensional space of 16384, i=1, 2, …, n.

Comparison of VMPE model with other works

We apply the VM-PE model upon the above-mentioned data set.

In order to assure our results, we list the observations for pairwise distance between Human and every one of the rest 19 species. As shown in Table 2, Columns 2~4 are the observations which are extracted from the pairwise distances matrices calculated by different approaches, i.e. Kolmogorov complexity (*Li et al.*, *2001*), OPT-SVD (*Yu & Huang*, *2012*) and our proposed VMPE model, respectively.

In general, the correlation degree between every two different results from each case is an effective way for comparing every two different approach. The higher correlation degree with the results of the existing approach and a newly developed one means that the latter has the same efficiency as the former.

For Kolmogorov complexity (*Li et al.*, *2001*), the calculated correlation coefficient value, between Column 2 and Column 4, is 0.95. Likewise, we can compute correlation coefficient value between Column 3 and Column 4, which is 0.8312. These phenomena indicate that our proposed VMPE model achieves the same effect as those representative approaches to the similarity analysis of genome sequences.

Phylogeny analysis of genomes via VMPE model

Generally, the steps for phylogenetic analysis among multiple genome sequences can be described as follows:

(1) Firstly, we can calculate all the corresponding projection feature vectors (FV) with different segmentation schemes for each genome sequence via VMPE;

(2) Secondly, selecting 'euclidean' distance metric, we calculate pair-wise distance matrix;
(3) Finally, we investigate the effect of dendrogram.

Fig. 2 illustrates the dendrogram derived from our proposed VMPE model, where these 20 species are separated clearly:

(a) Outgroup (Platypus, (Opossum, Wallaroo)) is far away from other clusters;
(b) Seven Primates are grouped together;
(c) Two Rodents (Mouse and Rat) stand at the same branch;
(d) The rest ones are clustered into a close group.

Meanwhile, these results are in agreement with both the evolutional facts and the conclusions in (*Li et al.*, *2001, Yu & Huang*, *2012*). Thus, it is shown that the proposed approach is effective in multiple genome sequences comparison. However, the result suggests the hypothesis of (Primates, (Rodents, Ferungulates)), which can also be found in Ref. (*Cao et al.*, *1998*), where there is a controversy on the hypothesis.

As a contrast, using MEGA software, we rebuild the Neighboring-Joining (NJ) tree through alignment-based method. As for the tree shown in Fig. 3, when compared to Fig. 2, it can be found that the proposed VMPE model yields similar results to that from the traditional approach.

Application upon another two genome datasets

To further verify the efficiency for our proposed VMPE model, we select another two genome datasets: 1) with 34 mammalian sequences, which has been also investigated by many works (*Huang et al.*, *2011, Yu et al.*, *2010, Yu*, *2013*); 2) with 16 longer sequences from a subfamily of Archaea, which has also been investigated by (*Qi et al.*, *2004*).

Larger dataset with 34 mammalian genome sequences

Likewise, according to procedure depicted, we obtain 34 feature vectors extracted from corresponding primary genome sequences via VMPE model. Moreover, these feature vectors have uniform dimension reduced to 34. Then, using these 34 vectors, we can calculate the pairwise distance matrix, through which we can construct the dendrogram for these 34 sequences. Fig. 4 illustrates the dendrogram based on VMPE model, while Fig. 5 shows the Neighboring-Joining (NJ) tree via alignment-based approach using MEGA software. Compare with Fig. 5, Fig. 4 illustrates that our proposed model produce a reasonable result similar to that from the traditional approach. Meanwhile, the clustering results of our model are also consistent with those in the representative published works (*Huang et al.*, *2011, Yu et al.*, *2010, Yu*, *2013*).

Longer dataset with 16 Archaea sequences

Even for a subfamily, where all the sequences are closer to each other, our proposed VMPE model still distinguishes them clear. The concise information the third dataset is listed in Table 3, while Fig. 6 shows the dendrogram.

DISCUSSION

In order to infer evolutional relationship among organism, we can use alignment-free approaches for estimating the evolutionary distances among multiple DNA/protein sequences and constructing phylogenetic trees. Most methods are based on *k*-mer patterns or transformation

from original sequences into numerical vectors. In general, such approaches are less accurate than those traditional phylogeny ones that are based on multiple sequence alignments (MSA). In k-mer based transformation, similarity analysis among multiple genome sequences may easily cause dimension curse, because it is a typical high dimension and small sample problem.

In this paper, our goal is to introduce a new methodology for the comparative genomics research community. As an alignment-free approach, our proposed VMPE model does not require any Human intervention and any evolution assumption. It is mathematically well-founded according to ICA-based (independent component analysis) distance preserving transformation which was strictly proved in theory. The latent application of our model is in the field of feature extraction with dimension reduced greatly. It works well when we use hierarchy ICA-based extractor to capture essential information to compare genome sequence within lower dimension space. Motivated by this approach, we proposed to use projected coordinates vector instead of the traditionally used *k*-mer based vectors directly to calculate the genetic distances among multiple sequences and to reconstruct phylogenetic trees.

Although a possible criticism for our approach is that it depends on the existing ICA technique, it is worth our while to stress that the dependence upon ICA-based approach is just to obtain the projected coordinates from multiple genome sequences, so that we can efficiently characterize the sequences within lower dimensional space. In fact, through projection, the proposed approach has a property of distances preserving, which achieves the reasonable results for the comparison of multiple genome sequences. Fig. 2 and Fig. 3 demonstrate that the dendrogram from our approach accords with the evolution tree obtained by traditional alignment-based. So do the results of Fig. 4 and Fig. 5.

However, the proposed model depends on an assumption that both local similarity and global similarity should be considered simultaneously for each genome sequence. Our preliminary experimental results have demonstrated the validity of the assumption. Moreover, our results have demonstrated that if the mixed vectors are projected in their independent-components-based coordinate system to extract the corresponding feature vectors within dimension-reduced space, the obtained lower-dimensional vectors are of great value to be applied into the fields of clustering and classification.

CONCLUSIONS

As an alignment free approach, our proposed VMPE model does not require any Human intervention and any evolution assumption. It is mathematically well-founded according to ICA-based distance preserving transformation which was strictly proved in theory. The latent application of our model is in the field of feature extraction with dimension reduced greatly. It works well when we use hierarchy ICA-based extractor to capture essential information to compare genome sequence within lower dimension space. Although a possible criticism for our approach is that it depends on alignment-based results, what is worth while stressing is that the dependence upon alignment is just to obtain the optimized segmentation scheme. In fact, through

optimization, the optimal segment number is about 3, which achieves the best performance on the comparison of genome.

However, the proposed model depends on an assumption that both local similarity and global similarity should be considered simultaneously for each genome sequence. Our preliminary experimental results have demonstrated the validity of the assumption. Moreover, our results have demonstrated that if the mixed vectors are projected in their independent-components-based coordinate system to extract the corresponding feature vectors within dimension-reduced space, the obtained lower-dimensional vectors are of great value to be applied into the fields of clustering and classification. In the future, we are planning to design a new output-controllable model to further improve the performance on similarity analysis or to extend its applications into other fields.

ACKNOWLEDGEMENTS

REFERENCES

Blaisdell BE. **1986**. A measure of the similarity of sets of sequences not requiring sequence alignment. *Proceedings of the National Academy of Sciences* **83(14)**:5155-5159 DOI 10.1073/pnas.83.14.5155.

Cao Y, Janke A, Waddell PJ, Westerman M, Takenaka O, Murata S, Okada N, Pääbo S, Hasegawa M. **1998**. Conflict among individual mitochondrial proteins in resolving the phylogeny of eutherian orders. *Journal of Molecular Evolution* **47(3)**:307-322 DOI 10.1007/PL00006389.

Chan RH, Chan TH, Yeung HM, Wang RW. **2012**. Composition vector method based on maximum entropy principle for sequence comparison. *IEEE/ACM Trans Comput Biol Bioinform* **9(1)**:79-87 DOI 10.1109/TCBB.2011.45.

Comon P. **1994**. Independent component analysis, a new concept? *Signal Processing* **36(3)**:287-314 DOI 10.1016/0165-1684(94)90029-9.

Cristea PD. **2002**. Conversion of nucleotide sequences into genomic signals. *Journal of Cellular and Molecular Medicine* **6(2)**:25 DOI 10.1111/j.1582-4934.2002.tb00196.x.

Dai Q, Liu X, YuhuaYao, Zhao F. **2011**. Numerical characteristics of word frequencies and their application to dissimilarity measure for sequence comparison. *Journal of Theoretical Biology* **276(1)**:174-180 DOI 10.1016/j.jtbi.2011.02.005.

Deng M, Yu C, Liang Q, He RL, Yau SS-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.g001.

Dong R, Zhu Z, Yin C, He RL, Yau SS. **2018**. A new method to cluster genomes based on cumulative Fourier power spectrum. *Gene* **673**:239-250 DOI 10.1016/j.gene.2018.06.042.

Fernandes F, Freitas AT, Almeida JS, Vinga S. **2009**. Entropic profiler - detection of conservation in genomes using information theory. *BMC Res Notes* **2**:72 DOI 10.1186/1756-0500-2-72.

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.

Huang DS, Zheng CH. **2006**. Independent component analysis-based penalized discriminant method for tumor classification using gene expression data. *Bioinformatics* **22(15)**:1855-1862 DOI 10.1093/bioinformatics/btl190.

Huang G, Zhou H, Li Y, Xu L. **2011**. Alignment-free comparison of genome sequences by a new numerical characterization. *Journal of Theoretical Biology* **281(1)**:107-112 DOI 10.1016/j.jtbi.2011.04.003.

Huang Y, Wang T. **2011**. Phylogenetic analysis of DNA sequences with a novel characteristic vector. *Journal of Mathematical Chemistry* **49**(**8**):1479-1492 DOI 10.1007/s10910-011-9811-x.

Huang Y, Yang L, Wang T. **2011**. Phylogenetic analysis of DNA sequences based on the generalized pseudo-amino acid composition. *Journal of Theoretical Biology* **269**(**1**):217-223 DOI 10.1016/j.jtbi.2010.10.027.

Jun S-R, Sims GE, Wu GA, Kim S-H. **2010**. Whole-proteome phylogeny of prokaryotes by feature frequency profiles：an alignment-free method with optimal feature resolution. *Proceedings of the National Academy of Sciences* **107**(**1**):6 DOI 10.1073/pnas.0913033107.

Li M, Badger JH, Chen X, Kwong S, Kearney P, Zhang H. **2001**. An information-based sequence distance and its application to whole mitochondrial genome phylogeny. *Bioinformatics* **17**(**2**):149-154 DOI 10.1093/bioinformatics/17.2.149.

Luczak BB, James BT, Girgis HZ. **2017**. A survey and evaluations of histogram-based statistics in alignment-free sequence comparison. *Brief Bioinform* (**1**):16 DOI 10.1093/bib/bbx161.

Mendizabal-Ruiz G, Roman-Godinez I, Torres-Ramos S, Salido-Ruiz RA, Velez-Perez H, Morales JA. **2018**. Genomic signal processing for DNA sequence clustering. *PeerJ* **6**(**3**):e4264 DOI 10.7717/peerj.4264.

Nguyen KD. **2012**. On the edge of web-based multiple sequence alignment services. *Tsing Hua Science and Technology* **17**(**6**):629-637 DOI 10.1109/TST.2012.6374364.

Qi J, Wang B, Hao B-I. **2004**. Whole proteome prokaryote phylogeny without sequence alignment a: k-string composition approach. *Journal of Molecular Evolution* **58**:1-11 DOI 10.1007/s00239-003-2493-7.

Sims GE, Jun S-R, Wu GA, Kim S-H. **2009**. Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. *Proceedings of the National Academy of Sciences* **106**(**8**): 2677-2682 DOI 10.1073/pnas.0813249106.

Sims GE, Jun SR, Wu GA, Kim SH. **2009**. Whole-genome phylogeny of mammals: Evolutionary information in genic and nongenic regions. *Proceedings of the National Academy of Sciences* **106**(**40**):17077-17082 DOI 10.1073/pnas.0909377106.

Sims GE, Kim S-H. **2011**. Whole-genome phylogeny of Escherichia colil/Shigella group by feature frequency profiles (FFPs). *Proceedings of the National Academy of Sciences* **108** (**20**):8329-8334 DOI 10.1073/pnas.

Stuart GW, Berry MW. **2004**. An SVD-based comparison of nine whole eukaryotic genomes supports a coelomate rather than ecdysozoan lineage. *BMC Bioinformatics* **5:**1-13 DOI 10.1186/1471-2105-5-204.

Stuart GW, Moffett K, Leader JJ. **2002**. A comprehensive vertebrate phylogeny using vector representations of protein sequences from whole genomes. *Molecular Biology Evolution* **19**(**4**):554-562 DOI .

Vinga S. **2014**. Information theory applications for biological sequence analysis. *Brief Bioinform* **15**(**3**):376-389 DOI 10.1093/bib/bbt068.

Vinga S, Almeida J. **2003**. Alignment-free sequence comparison--a review. *Bioinformatics* **19**(**4**):513-523 DOI 10.1093/bioinformatics/btg005.

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**(**22**):4125-4132 DOI 10.1093/bioinformatics/bti658.

Yang L, Zhang X, Zhu H. **2012**. Alignment free comparison: similarity distribution between the DNA primary sequences based on the shortest absent word. *Journal of Theoretical Biology* **295**:125-131 DOI 10.1016/j.jtbi.2011.11.021.

Yeredor A. **2002**. Non-orthogonal joint diagonalization in the least-squares sense with application in blind source separation. *IEEE Transactions on Signal Processing* **50**(**7**):1545-1553 DOI 10.1109/TSP.2002.1011195.

Yu C, Liang Q, Yin C, He RL, Yau SST. **2010**. A novel construction of genome space with biological geometry. *DNA Research* **17**(**3**):155-168 DOI 10.1093/dnares/dsq008.

Yu H-J. **2013**. Segmented K-mer and its application on similarity analysis of mitochondrial genome sequences. *Gene* **518**(**2**):419-424 DOI 10.1016/j.gene.2012.12.079.

Yu H-J, Huang D-S. **2012**. Novel graphical representation of genome sequence and its applications in similarity analysis. *Physica A: Statistical Mechanics and its Applications* **391**(**23**):6128-6136 DOI 10.1016/j.physa.2012.07.020.

Table 1(on next page)

Summary information for the 20 eutherians species

1 **Table 1.** Summary information for the 20 eutherians species
2

Accession no.	Species	Length
V00662	Human	16,569
D38116	Pigmy chimpanzee	16,563
D38113	Common chimpanzee	16,554
D38114	Gorilla	16,364
D38115	Bornean orangutan	16,389
X99256	Gibbon	16,472
Y18001	Baboon	16,521
X79547	Horse	16,660
Y07726	White rhinoceros	16,832
X63726	Harbor seal	16,826
X72004	Gray seal	16,797
U20753	Cat	17,009
X61145	Fine Whale	16,398
X72204	Blue Whale	16,402
V00654	Cow	16,338
X14848	Norway rat	16,300
V00711	Mouse	16,295
Z29573	Opossum	17,084
Y10524	Wallaroo	16,896
X83427	Platypus	17,019

3

Table 2(on next page)

Comparison of performance values with other two published works for the 20 eutherians species via the pairwise distances among Human and the rest 19 ones

Table 2. Comparison of performance values with other two published works for the 20 eutherians species via the pairwise distances among Human and the rest 19 ones

Human Vs. rest species	Kolmogorov complexity (*Li, Badger, Chen, Kwong, Kearney & Zhang, 2001*)	OPT-SVD (*Yu & Huang, 2012*)	VMPE ($\times 10^{-2}$)
P.Chim.	0.654234	0.059285712	0.414459
C.Chim.	0.657387	0.041232607	0.413945
Gorilla	0.732325	0.041462402	0.442228
Orang.	0.847139	0.068655059	0.486724
Gibbon	0.880203	0.065596946	0.496526
Baboon	0.841775	0.070727095	0.533934
Horse	0.971558	0.144379093	0.555873
W.Rhin.	0.973694	0.168378304	0.557738
H.Seal	0.974737	0.211514019	0.578971
G.Seal	0.97576	0.207692966	0.579792
Cat	0.977328	0.276304513	0.577581
F.Whale	0.980493	0.209182609	0.562304
B.Whale	0.976034	0.197653729	0.557521
Cow	0.97362	0.283610943	0.565623
Rat	0.981715	0.276197913	0.569249
Mouse	0.9804	0.372617581	0.585258
Oposs.	0.986243	0.54735028	0.626082
Walla.	0.985926	0.258992019	0.569803
Platypus	0.988041	0.445501566	0.612848

Li M, Badger JH, Chen X, Kwong S, Kearney P, Zhang H. **2001**. An information-based sequence distance and its application to whole mitochondrial genome phylogeny, *Bioinformatics*, **17**(**2**):149-154 DOI 10.1093/bioinformatics/17.2.149.

Yu H-J, Huang D-S. **2012**. Novel graphical representation of genome sequence and its applications in similarity analysis, *Physica A: Statistical Mechanics and its Applications*, **391**(**23**):6128-6136 DOI 10.1016/j.physa.2012.07.020.

Table 3(on next page)

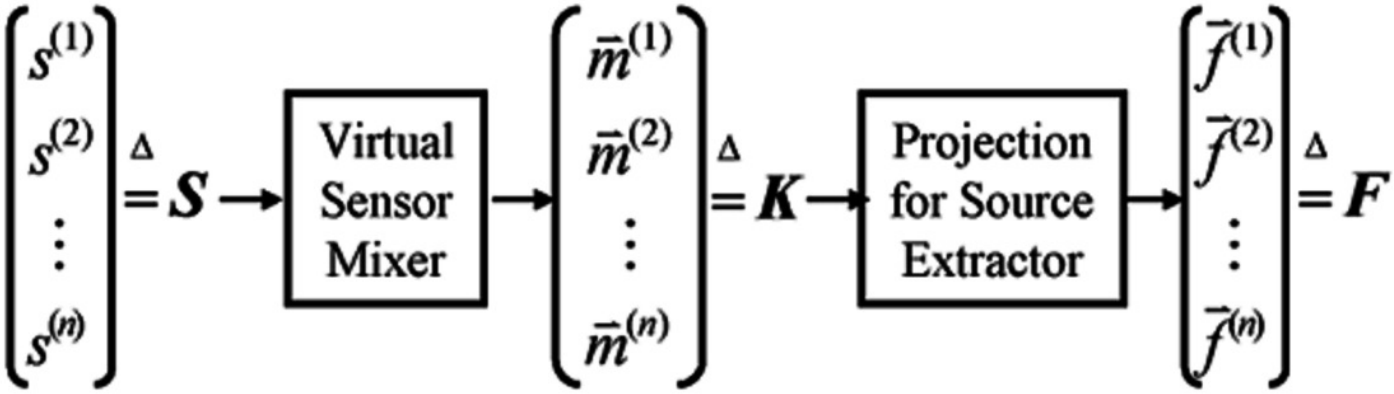
16 longer sequences of Archaea with names, abbreviations, and NCBI accession numbers

Table 3. 16 longer sequences of Archaea with names, abbreviations, and NCBI accession numbers

Species/strain	Abbrev.	Accession No.	Length (nt)
Pyrobaculum aerophilum	01Pyrae	NC_003364	2222430
Aeropyrum pernix Kl	02Aerpe	NC_000854	1669696
Sulfolobus solfataricus	03Sulso	NC_002754	2992245
Sulfolobus tokodaii	04Sulto	NC_003106	2694756
Methanobacterium thermoautotrophicus	05Metth	NC_000916	1751377
Methanococcus jannaschii	06Metja	NC_000909	1664970
Methanosarcina acetivorans strain C2A	07Metac	NC_003552	5751492
Methanosarcina mazei Goel	08Metma	NC_003901	4096345
Halobacterium sp. NRC-1	09Halsp	NC_002607	2014239
Thermoplasma acidophilum	10Theac	NC_002578	1564906
Thermoplasma volcanium	11Thevo	NC_002689	1584804
Pyrococcus abyssi	12Pyrab	NC_000868	1765118
Pyrococcus furiosus	13Pyrfu	NC_003413	1908256
Pyrococcus horikoshii	14Pyrho	NC_000961	1738505
Archaeoglobus fulgidus	15Arcfu	NC_000917	2178400
Methanopyrus kandleri AV19	16Metka	NC_003551	1694969

Figure 1

Scheme for the monolayer ICA upon k-mers of corresponding multiple sequences



PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27333v1 | CC BY 4.0 Open Access | rec: 8 Nov 2018, publ: 8 Nov 2018

Figure 2

The dendrogram based on segmented 7-mer using 20 mitochondrial genome sequences of eutherian 366 species

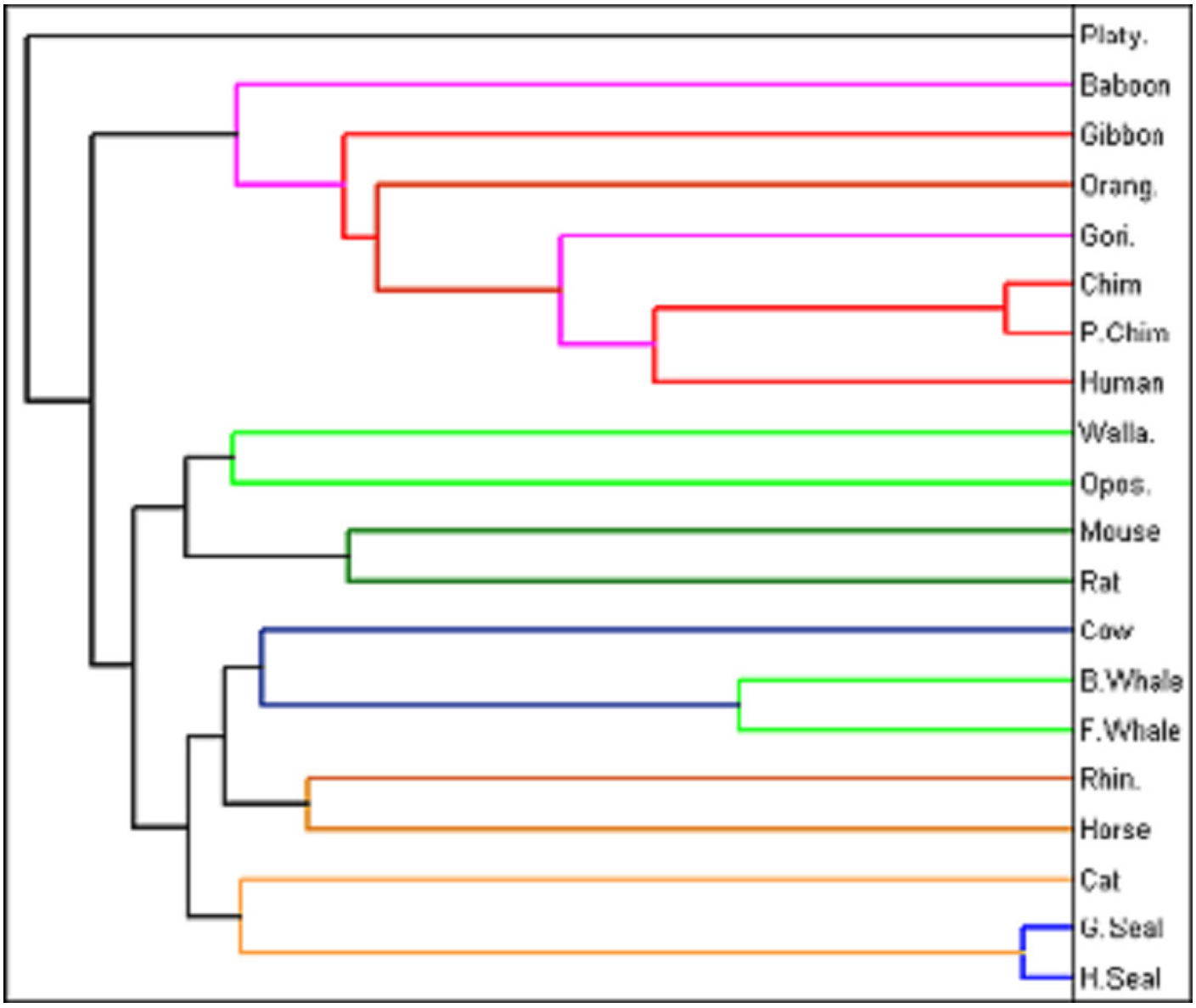


Figure 3

The Neighboring-Joining (NJ) tree of the eutherians constructed from the mitochondrial genome 370 sequences of the 20 species via traditional alignment-based approach using MEGA software

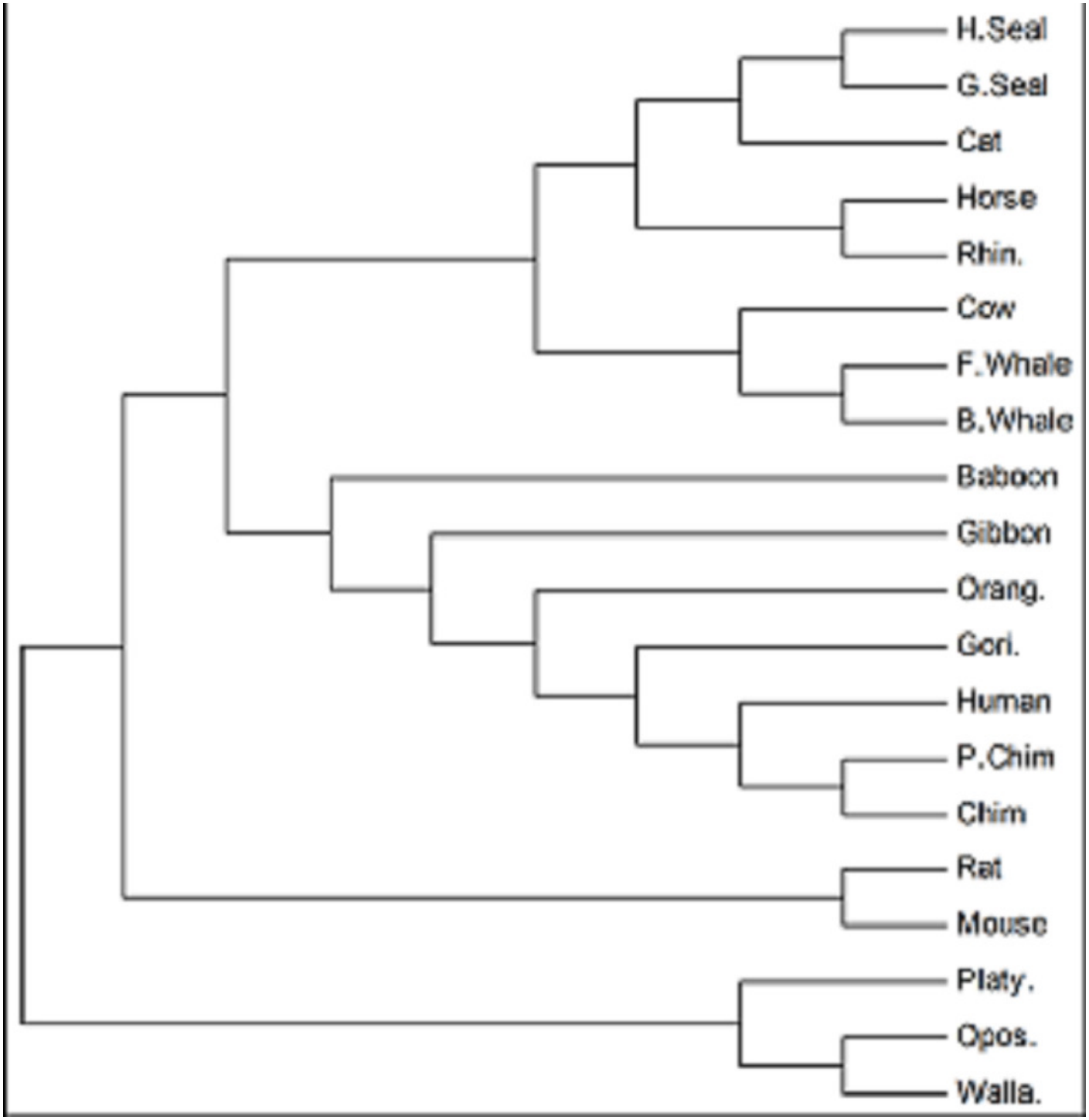


Figure 4

The dendrogram based on VMPE model using 34 mitochondrial genome sequences from mammalian 400 species

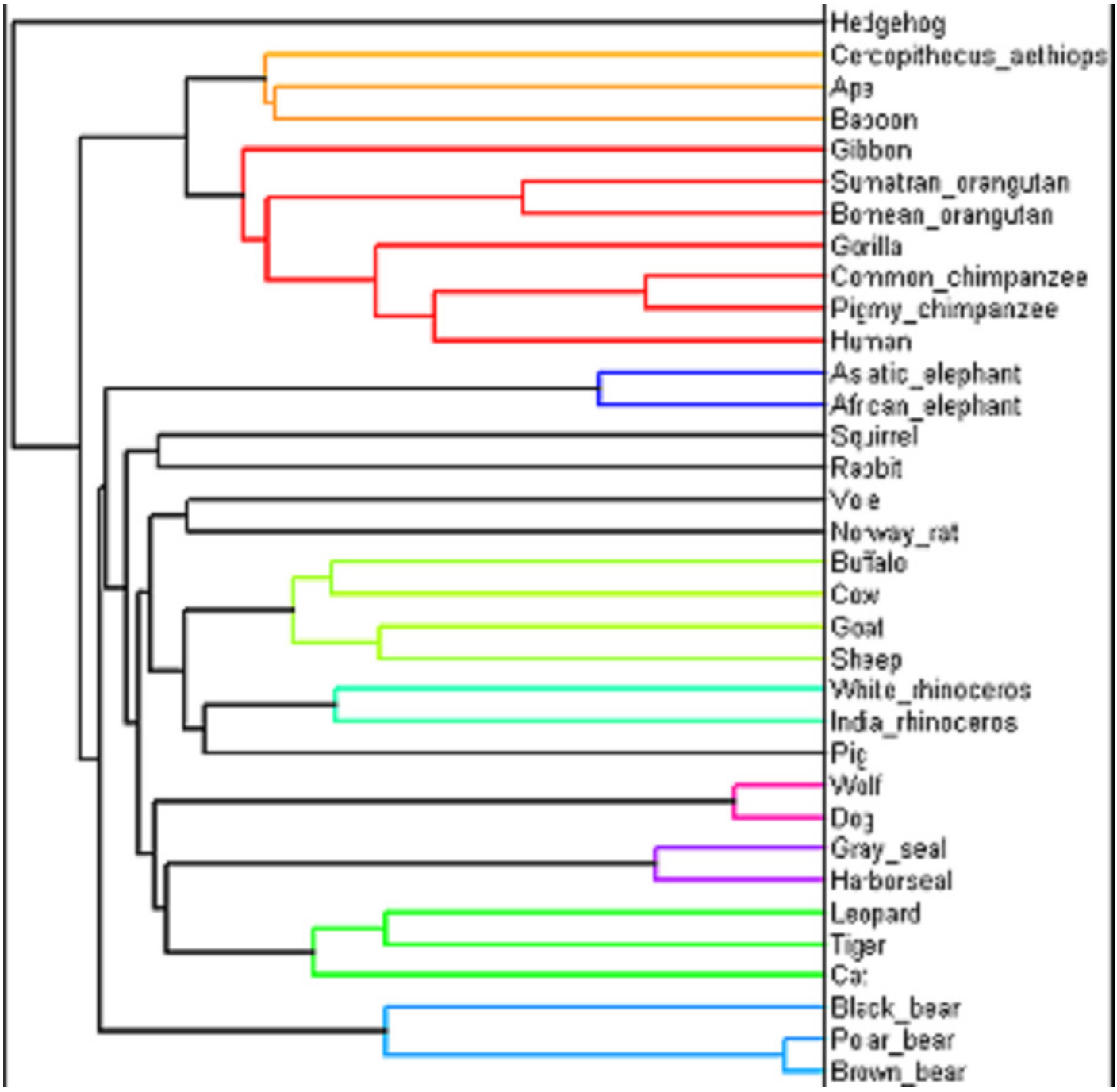
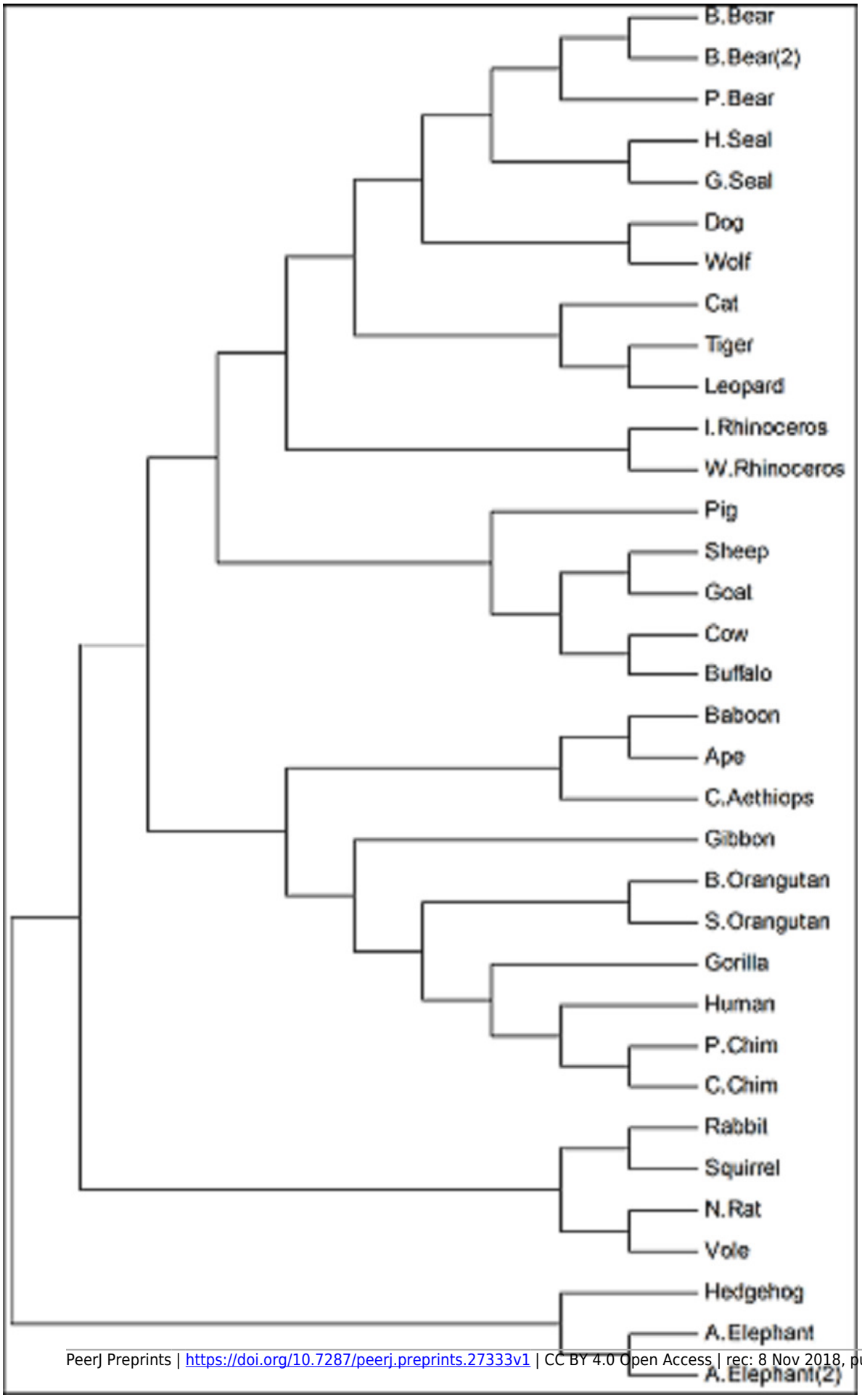


Figure 5

The Neighboring-Joining (NJ) tree of the mammals constructed from the mitochondrial genome 405 sequences of the 34 species via traditional alignment-based approach using MEGA software

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27333v1 | CC BY 4.0 Open Access | rec: 8 Nov 2018, publ: 8 Nov 2018



PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27333v1 | CC BY 4.0 Open Access | rec: 8 Nov 2018, publ: 8 Nov 2018

Figure 6

The dendrogram based on VMPE model using 16 Archaea sequences