Amino Acids (2012) 42:1867–1877
DOI 10.1007/s00726-011-0906-2

ORIGINAL ARTICLE

Using Markov model to improve word normalization algorithm for biological sequence comparison

Qi Dai · Xiaoqing Liu · Yuhua Yao · Fukun Zhao

Received: 13 November 2010 / Accepted: 29 March 2011 / Published online: 20 April 2011

Abstract There are two crucial problems with statistical measures for sequence comparison: overlapping structures and background information of words in biological sequences. Word normalization in improved composition vector method took into account these problems and achieved better performance in evolutionary analysis. The word normalization is desirable, but not sufficient, because it assumes that the four bases A, C, T, and G occur randomly with equal chance. This paper proposed an improved word normalization which uses Markov model to estimate exact *k*-word distribution according to observed biological sequence and thus has the ability to adjust the background information of the *k*-word frequencies in biological sequences. The improved word normalization was tested with three experiments and compared with the existing word normalization. The experiment results confirm that the improved word normalization using Markov model to estimate the exact *k*-word distribution in biological sequences is more efficient.

Keywords Markov model · Word normalization · Sequence comparison · Classification · Phylogenetic analysis

Electronic supplementary material The online version of this article (doi:10.1007/s00726-011-0906-2) contains supplementary material, which is available to authorized users.

Q. Dai (✉) · Y. Yao · F. Zhao
College of Life Sciences, Zhejiang Sci-Tech University, Hangzhou 310018, People's Republic of China
e-mail: daiailiu2004@yahoo.com.cn

X. Liu
School of Science, Hangzhou Dianzi University, Hangzhou 310018, People's Republic of China

Introduction

Sequence comparison helps us to find both similarities and differences among biological sequences, and assign the function of those sequences (Mitrophanov and Borodovsky 2006; Altschul et al. 1997; Dai et al. 2008; Pham and Zuegg 2004; Pham 2007; Felsenstein 1996; Huelsenbeck and Ronquist 2001; Komatsu et al. 2001; Kumar et al. 2004; Li et al. 2001; Otu and Sayood 2003; Ronquist and Huelsenbeck 2003; Waddell et al. 2001; Mohseni-Zadeh et al. 2004; Pipenbacher et al. 2002; Yao et al. 2008; Yang et al. 2010). For example, to identify genes and functionally related regulatory sequences in newly sequenced genomes, one must discern functional similarity among candidate subsequences based on their similarity/dissimilarity. Nowadays, numerous tools for sequence comparisons were developed, and were classified into two categories: alignment-based and alignment-free. Alignment is a typical method for sequence comparison, its comprehensive reviews can be found in Durbin et al. (1998) and Waterman (1995). However, the alignment-based methods have both fundamental and computational limitations (Pham and Zuegg 2004; Vinga and Almeida 2003). For example, these methods cannot deal with changes, such as chromosome reversal, gene translocation and the large data of biological sequences. Consequently, considerable efforts have been made to seek for alternative, i.e., alignment-free, methods for sequence comparison.

Word-based statistical model is one of the most well-developed alignment-free methods, which were recently reviewed by Vinga and Almeida (2003). Among the word-based models, each sequence is mapped into an *m*-dimensional vector according to its *k*-word frequencies. The similarity score among sequences represented in vector spaces is then defined by measures, such as Euclidean

distance (Blaisdell 1986), Cosine distance (Stuart et al. 2002), Mahalanobis distance (Wu et al. 1997) and Kullback–Leibler discrepancy (Wu et al. 2001) between their corresponding vectors. Although the word-based statistical measures have been successfully applied to sequence comparison, it has to be noted that the standard approach for calculating the word frequencies has its limitation. Because it gives emphasis to the total occurrences of the words and ignores the overlapping structure of the words, which may cause information loss.

If the words occurring in biological sequence are estimative probabilities rather than the frequency, they are more readily optimized by objective mathematical models. This enables building more complex, biologically realistic models with large numbers of parameters, such as Markov model (Pham and Zuegg 2004; Hao and Qi 2004; Wu et al. 2006), mix model such as Markov model plus k-word distributions (Dai et al. 2008; Kantorovitz et al. 2007), and Bernoulli model assuming a known word distribution (Lu et al. 2008). Although the more complex models in biological sequence comparison are general improvements over the traditional word-based models (Blaisdell 1986; Wu et al. 1997, 2001; Stuart et al. 2002), some problems in developing statistical models and estimating the parameters of the complex models have impeded the development and adoption of these or other more complex models.

Recently, Lu et al. (2008) proposed an improved composition vector (ICV) method that takes into consideration the above problems and achieves better performance in sequence comparison. The word normalization in improved composition vector is desirable, but not sufficient, because much effort of the word normalization aims to find better ways of utilizing evolution information. It is necessary to build more widely used and efficient word normalization to improve sequence comparison. This paper presents an improved word normalization for sequence comparison, which has ability to estimate the word distribution based on the observed biological sequences. The contents can be summarized as follows:

1. An improved word normalization for sequence comparison was proposed, in which expectation and variance of the frequencies of the k-word were estimated according to the observed biological sequences under Markov model. The improved word normalization takes into consideration the overlapping occurrences and has the ability to adjust the background information of the k-word frequencies in biological sequences.
2. Effectiveness of the improved word normalization were evaluated by extensive experiments, including discrimination between functionally related regulatory sequences and unrelated sequences, intron and exon. Moreover, their effectiveness is compared with the existing word normalization. Through the experiments, we want to address effectiveness of the proposed improved word normalization, and whether the improved word normalization improve the performance of the word-based measures for phylogenetic analysis.

Materials and methods

Word statistics

A biological sequence is interpreted as a succession of symbols and a k-word is a series of k consecutive letters in a sequence. The k-word statistical analysis consists of counting occurrences of k-words in a given sequence. For a sequence $s = s_1 s_1 \ldots$, the count of a k-word $w_k = w_{k,1} w_{k,2} \ \ldots \ w_{k,k}$, denoted by $c(w_k)$, is the number of occurrences of the word w_k in the sequence s. The standard approach for counting k-words in a sequence of length n is to use a sliding window of length k, shifting the frame one base at a time from position 1 to $n - k + 1$. In this method, k-words are allowed to overlap in the sequence. In this way, a sequence can be represented by an m-dimensional vector C_k^s made up of k-word counts

$$C_k^s = \left(c(w_k^1), c(w_k^2), \ldots, c(w_k^m)\right), \tag{1}$$

where m is the total number of all possible k-words. The vector of k-word frequencies $f(w_k)$, denoted by F_k^s, can be calculated by

$$\begin{aligned} F_k^s &= \left(f(w_k^1), f(w_k^2), \ldots, f(w_k^m)\right) \\ &= \left(\frac{c(w_k^1)}{n-k+1}, \frac{c(w_k^2)}{n-k+1}, \ldots, \frac{c(w_k^m)}{n-k+1}\right). \end{aligned} \tag{2}$$

For example, consider the DNA sequence $s = \mathrm{AAAGGA}$, we can obtain the vectors made up of 2-word counts and frequencies

$$\begin{aligned} C_2^s &= \left(c(\mathrm{AA}), c(\mathrm{AG}), c(\mathrm{GG}), c(\mathrm{GA})\right) = (2, 1, 1, 1), \\ F_2^s &= \left(f(\mathrm{AA}), f(\mathrm{AG}), f(\mathrm{GG}), f(\mathrm{GA})\right) = (0.4, 0.2, 0.2, 0.2). \end{aligned}$$

Previous word normalization

The k-word counts' vector C_k^s reflects both random mutation and selection, and the random background needs to be normalized to represent genetic information contributed by natural selection. Based on this idea, composition vector (CV) was proposed and has been used with minor modifications for phylogenetic studies of prokaryotes and viruses (Hao and Qi 2004; Wu et al. 2006). The formula for composition vector (CV) of the observed frequency of the k-word, $\mathrm{CV}(w_{k,1} \ldots w_{k,k})$, in the sequence s is defined as below:

$$\mathrm{CV}(w_{k,1}\ldots w_{k,k}) = \frac{c(w_{k,1}\ldots w_{k,k}) - c^0(w_{k,1}\ldots w_{k,k})}{c^0(w_{k,1}\ldots w_{k,k})} \tag{3}$$

where

$$c^0(w_{k,1}\ldots w_{k,k}) = \frac{c(w_{k,1}\ldots w_{k,k-1})c(w_{k,2}\ldots w_{k,k})}{c(w_{k,2}\ldots w_{k,k-1})} \times \frac{(n-k+1)(n-k+3)}{(n-k+2)^2}$$

for $k \geq 3$.

Lu et al. (2008) have found two problems associated with composition vector methods: (a) there is a positive correlation between the observed count $c(w_{k,1}\ldots w_{k,k})$ and the estimated expected count $c^0(w_{k,1}\ldots w_{k,k})$, and (b) a square root needs to be applied to the denominator. Without such an operation, the normalized count tends to be over-standardized. To overcome the problems, Lu constructed an improved CV (ICV) of all k-word computed as follows:

$$\mathrm{ICV}(w_{k,1}\ldots w_{k,k}) = \frac{c(w_{k,1}\ldots w_{k,k}) - \mathbb{E}[c(w_{k,1}\ldots w_{k,k})]}{\sqrt{\mathbb{V}\mathrm{ar}[c(w_{k,1}\ldots w_{k,k})]}}, \tag{4}$$

where

$$\mathbb{E}[c(w_{k,1}\ldots w_{k,k})] = \frac{n-k+1}{4^k},$$

$$\mathbb{V}\mathrm{ar}[c(w_{k,1}\ldots w_{k,k})] = \frac{n-k+1}{4^k}\left(1-\frac{1}{4^k}\right) - \frac{2}{4^{2k}}(k-1)\left(n-\frac{3}{2}k+1\right) + \frac{2}{4^k}\sum_{t=1}^{k-1}(n-k+1-t)\frac{J_t}{4^t},$$

J_t is an indicator function, equal to 1 if $w_{k,1}\ldots w_{k,t} = w_{k,t+1}\ldots w_{k,k}$ and equal to 0 otherwise, for $t = 1, 2, \ldots, k-1$.

Improved word normalization based on Markov model

ICV method assumes that four bases A, C, T, and G occur randomly with equal chance and derives the expected and variance of a k-word count in a given sequence. In other words, the word distribution is assumed to be known a priori. However, in most cases, the word distribution is usually unknown, and therefore the ICV method is very limited.

This paper presents an improved word normalization in which the expectation and variance of the k-word frequencies is estimated with Markov model. Similar to the work (1999), Reinert et al. (2000), Schbath (2000), we used definitions and properties related to occurrences of the k-word $w_k = w_{k,1}w_{k,2}\ldots w_{k,k}$ in the biological sequence $s = s_1s_1\ldots$. In the remainder of this section, we consider Markov model with order 1 for the sequence s, generalizations to high order can be deduced similarly.

The position of an occurrence of w_k is defined by the position of its first letter $w_{k,1}$. The random indicator $Y_i(w_k)$ of an occurrence of w_k at position i, $1 \leq i \leq n-k+1$, in s is defined by

$$Y_i(w_k) = \begin{cases} 1 & \text{if}(s_i, s_{i+1}, \ldots, s_{i+k-1}) = (w_{k,1}, w_{k,2}, \ldots, w_{k,k}), \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

$Y_i(w_k)$ is a random Bernoulli variable with parameter given under Markov model (M) by

$$\mathbb{P}(Y_i(w_k) = 1|\mathrm{M}) = \mathbb{P}(s_i = w_{k,1}, \ldots, s_{i+k-1} = w_{k,k}) = \pi(w_{k,1})\prod_{j=2}^{k} p(w_{k,j-1}, w_{k,j}), \tag{6}$$

where $\mathbb{P}(Y_i(w_k) = 1)$ denotes the probability of the word w_k appearing at the position i under Markov model (M), $p(w_{k,i}, w_{k,j})$ is transition probability going from the state $w_{k,i}$ to another state $w_{k,j}$, and $\pi(w_{k,1}$ denotes the initial state distribution of the state $\pi(w_{k,1})$. For convenience, let $\mu(w_k)$ denotes the probability for the word w_k to appear at a given position in the sequence. Because the $Y_i(w_k)$ is a Bernoulli variable, its expectation $\mathbb{E}[Y_i(w_k|\mathrm{M}]$ and variance $\mathbb{V}[Y_i(w_k)|\mathrm{M}]$ under Markov model (M) can be calculated as follows:

$$\mathbb{E}[Y_i(w_k|\mathrm{M}] = \mu(w_k), \mathbb{V}[Y_i(w_k)|\mathrm{M}] = \mu(w_k)(1-\mu(w_k)). \tag{7}$$

So the expectation $\mathbb{E}[f(w_k)|\mathrm{M}]$ of the frequencies of the word w_k under Markov model(M) is

$$\mathbb{E}[f(w_k)|\mathrm{M}] = \frac{\mathbb{E}[c(w_k)|\mathrm{M}]}{n-k+1} = \frac{\sum_{i=1}^{n-k+1}\mathbb{E}[Y_i(w_k)|\mathrm{M}]}{n-k+1}. \tag{8}$$

As pointed out by Reinert (2000), Robin (1999) and Schbath (2000), the variables $Y_i(w_k)$ and $Y_{i+d}(w_k)$ are not independent under Markov model. Therefore, we can calculate the covariance of $Y_i(w_k)$ and $Y_{i+d}(w_k)$, denoted by $\mathbb{C}\mathrm{ov}[Y_i(w), Y_{i+d}(w)|\mathrm{M}]$, as follows:

$$\mathbb{C}\mathrm{ov}[Y_i(w), Y_{i+d}(w)|\mathrm{M}] = \begin{cases} \mu(w_k)\varepsilon_{k-d}(w_k)\prod_{j=k-d+1}^{k} p(w_{k,j-1}, w_{k,j}) - \mu(w_k)^2 & \text{if } 1 \leq d \leq k, \\ \mu(w_k)^2\left(\frac{p^{d-k+1}(w_{k,k}, w_{k,1})}{\pi(w_{k,1})} - 1\right) & \text{if } d \geq k. \end{cases} \tag{9}$$

where

$$\varepsilon_m(w_k) = \begin{cases} 1 & \text{if } (w_{k,k-m+1}, \ldots, w_{k,k}) = (w_{k,1}, \ldots, w_{k,m}), \\ 0 & \text{otherwise} \end{cases}$$

We then obtain

$$\begin{aligned} \mathbb{V}[f(w_k)|\mathrm{M}] &= \frac{\sum_{i=1}^{n-k+1} \mathbb{V}[Y_i(w_k)|\mathrm{M}] + \sum_{\mathrm{i}=1}^{\mathrm{n-k+1}} \sum_{\mathrm{j}=\mathrm{i}+1}^{\mathrm{n-k+1}} \mathbb{C}\mathrm{ov}[\mathrm{Y_i(w_k)}, \mathrm{Y_j(w_k)}|\mathrm{M}]}{(n-k+1)^2} \\ &= \Bigg((n-k+1)\mu(w_k)(1-\mu(w_k)) + 2\sum_{d=1}^{k-1}(n-d-k+1)\mu(w_k)\Big(\varepsilon_{k-d}(w_k) \prod_{j=k-d+1}^{k} p(w_{k,j-1}, w_{k,j}) - \mu(w_k)\Big) \\ &\quad + 2\mu(w_k)^2 \sum_{t=1}^{n-2k+1} (n-2k-t+2)\Big(\frac{p^{d-k+1}(w_{k,k}, w_{k,1})}{\pi(w_{k,1})} - 1\Big)\Bigg) \Big/ (n-k+1)^2. \end{aligned} \tag{10}$$

where $\mathbb{V}[f(w_k)|\mathrm{M}]$ is the variance of the k-word frequencies under Markov model (M).

Similar to the work of Lu et al. (2008), we normalized the frequencies of the k-words under Markov model (M), denoted by NF, as follows:

$$\mathrm{NF}(w_{k,1}\ldots w_{k,k}) = \frac{f(w_{k,1}\ldots w_{k,k}) - \mathbb{E}[f(w_{k,1}\ldots w_{k,k})|\mathrm{M}]}{\sqrt{\mathbb{V}\mathrm{ar}[f(w_{k,1}\ldots w_{k,k})|\mathrm{M}]}}. \tag{11}$$

Note that a word is called highly expressed if its observed frequency is more than its expected frequency, and called low expressed otherwise. In this sense, the normalization of k-word frequencies NF(w_k) measures a level of expression—large value of NF(w_k) corresponds to low expression of the word w_k, and small value of NF(w_k) corresponds to high expression of the word w_k.

The only difference between the ICV method and the proposed improved normalization function NF is the calculation of the expectation and variance of the word frequencies. The ICV method assumes that the four bases A, C, T, and G occur randomly with equal chance and derives the expectation and variance in a given sequence based on this simple assumption. In contrast, the proposed improved normalization function NF estimates the word expectation and variance according to the observed biological sequences with help of Markov model.

Evaluation methods

ROC analysis has been widely used in signal detection and classification problems (Egan 1975). This approach is employed in binary classification of continuous data, usually categorized as positive (1) or negative (0) cases. The classification accuracy can be measured by plotting, for different threshold values, the number of true positives (TP) and 1-specificity, encountered for each threshold, where

$$\begin{aligned} \text{Sensitivity} &= \frac{\text{True positives}}{\text{Positives}} = \frac{\text{TP}}{\text{TP+FN}}, \\ \text{Specificity} &= \frac{\text{True negatives}}{\text{Negatives}} = \frac{\text{TN}}{\text{TN+FP}}, \\ 1 - \text{Specificity} &= \frac{\text{FP}}{\text{TN+FP}}. \end{aligned} \tag{12}$$

ROC curve is a graphical plot of sensitivity versus (1-specificity) for different threshold values. The area under an ROC curve (AUC) is a widely employed parameter to quantify the quality of a classificator because it is a threshold independent performance measure and is closely related to the Wilcoxon signed-rank test (Bradley 1997). For a perfect classifier, the AUC is 1 and for a random classifier the AUC is 0.5.

Results and discussion

Evaluation on functionally related regulatory sequences

The improved word normalization is tested to evaluate if functionally or evolutionarily related sequence pairs are scored better than unrelated pairs of sequences randomly chosen from the genome. To assess the performance on functionally related sequences, we construct data sets as follows. The set of *cis-regulatory* modules (CRM), known to regulate expression in the same tissue, is taken as the 'positive' set. The set of randomly chosen non-coding sequences, with lengths and the total number matching the CRMs, is taken as the 'negative' set. The following seven data sets are chosen to analyze: FLY BLASTODERM (82 CRMs with expression in the blastoderm-stage embryo of

the fruitfly, *D. melanogaster*); FLY PNS (23 CRMs (average length 998 bp) driving expression in the peripheral nervous system in the fruitfly); FLY TRACHEAL [9 CRMs (average length 1,220 bp] involved in regulation of the tracheal system in the fruitfly); FLY EYE [17 CRMs (average length 894 bp) expressing in the *Drosophila* eye]; HUMAN MUSCLE [28 human CRMs (average length 450) regulating muscle specific gene expression]; HUMAN LIVER (9 CRMs (average length 201) driving expression specific to the human liver); HUMAN HBB [17 CRMs (average length 453) regulating the HBB complex]. They are well studied in Dai et al. (2008), Gallo et al. (2006), Kantorovitz et al. (2007).

Each pair of sequences in the positive set is compared, and so is each pair in the negative set. The evaluation is based on a binary classification of each sequence pair, where 1 corresponds to the pairs from positive set, 0 corresponds to the pairs from negative set. Let N be the number of sequences in the positive set, all the pairs constitute a vector of length $2\binom{N}{2}$,which is used as prediction. In addition, we can get a vector of length $2\binom{N}{2}$, consisting of 1 and 0 as class labels. A perfect measure would completely separate negative from positive set. Of

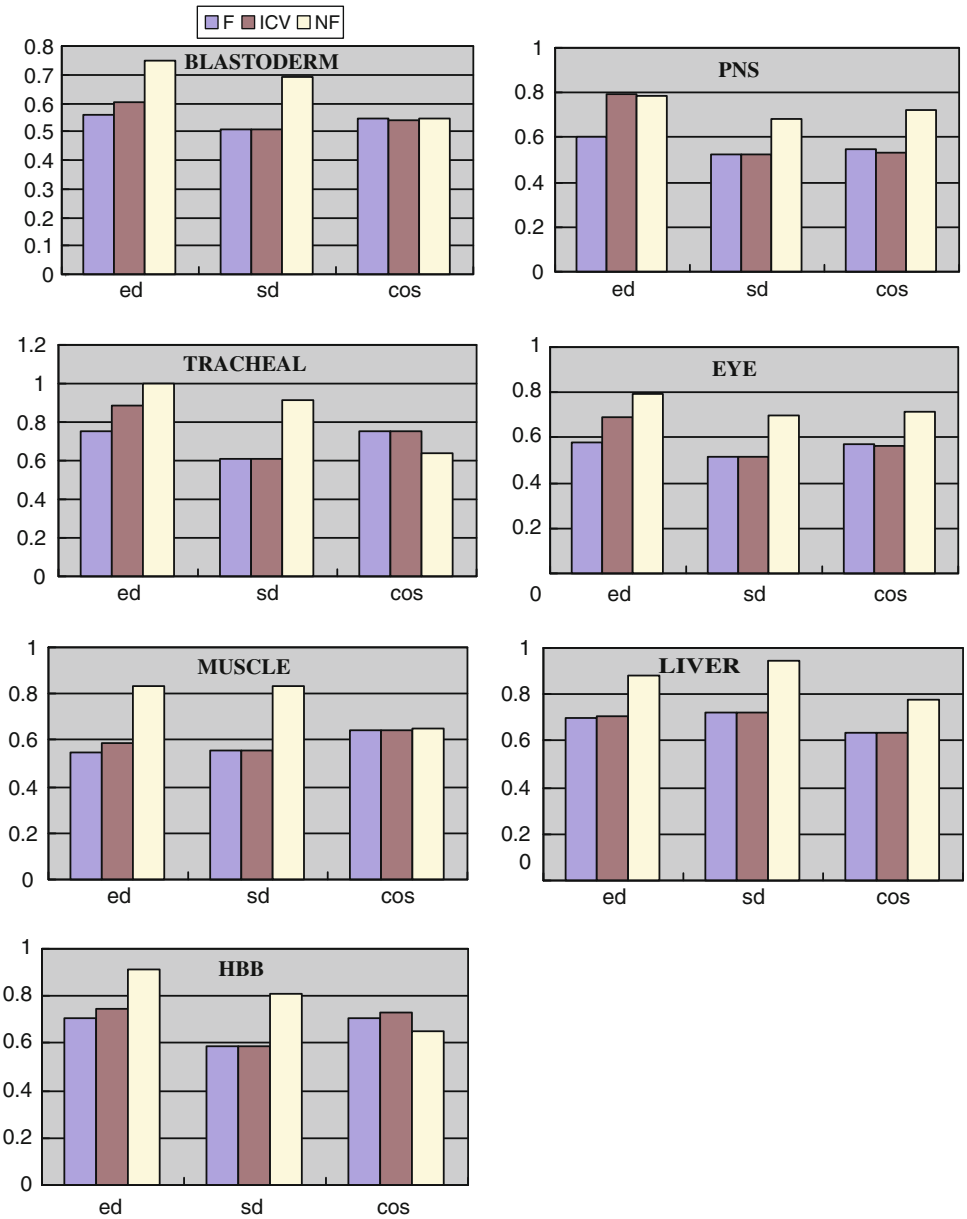


Fig. 1 Comparison of AUCs of three word-base measures based on *k*-word frequencies (*F*), improved composition vector (ICV) and word normalization (NF) for seven data sets

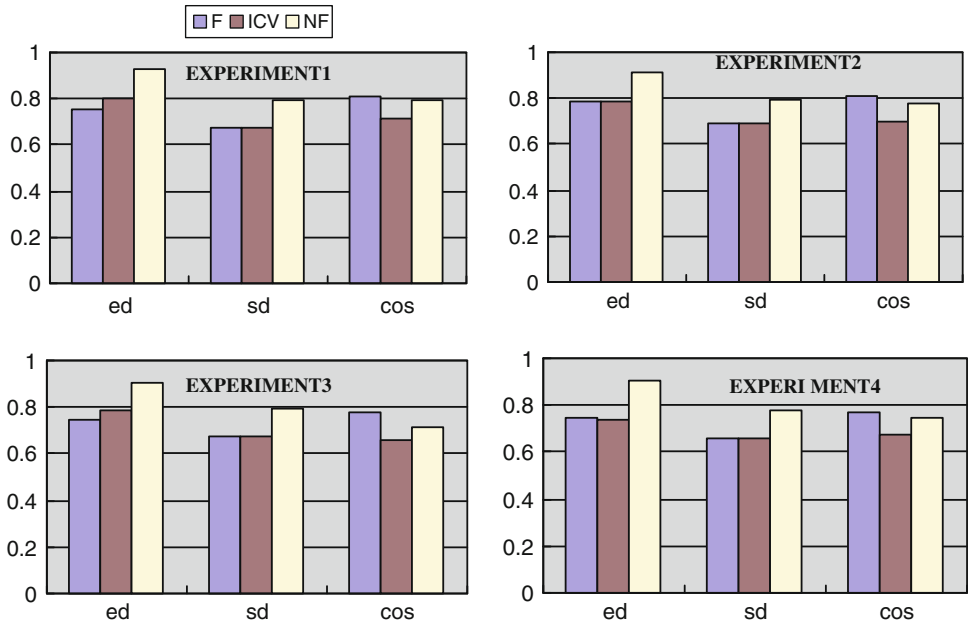


Fig. 2 Comparison of AUCs of three word-base measures based on *k*-word frequencies (*F*), improved composition vector (ICV) and word normalization (NF) for four data sets

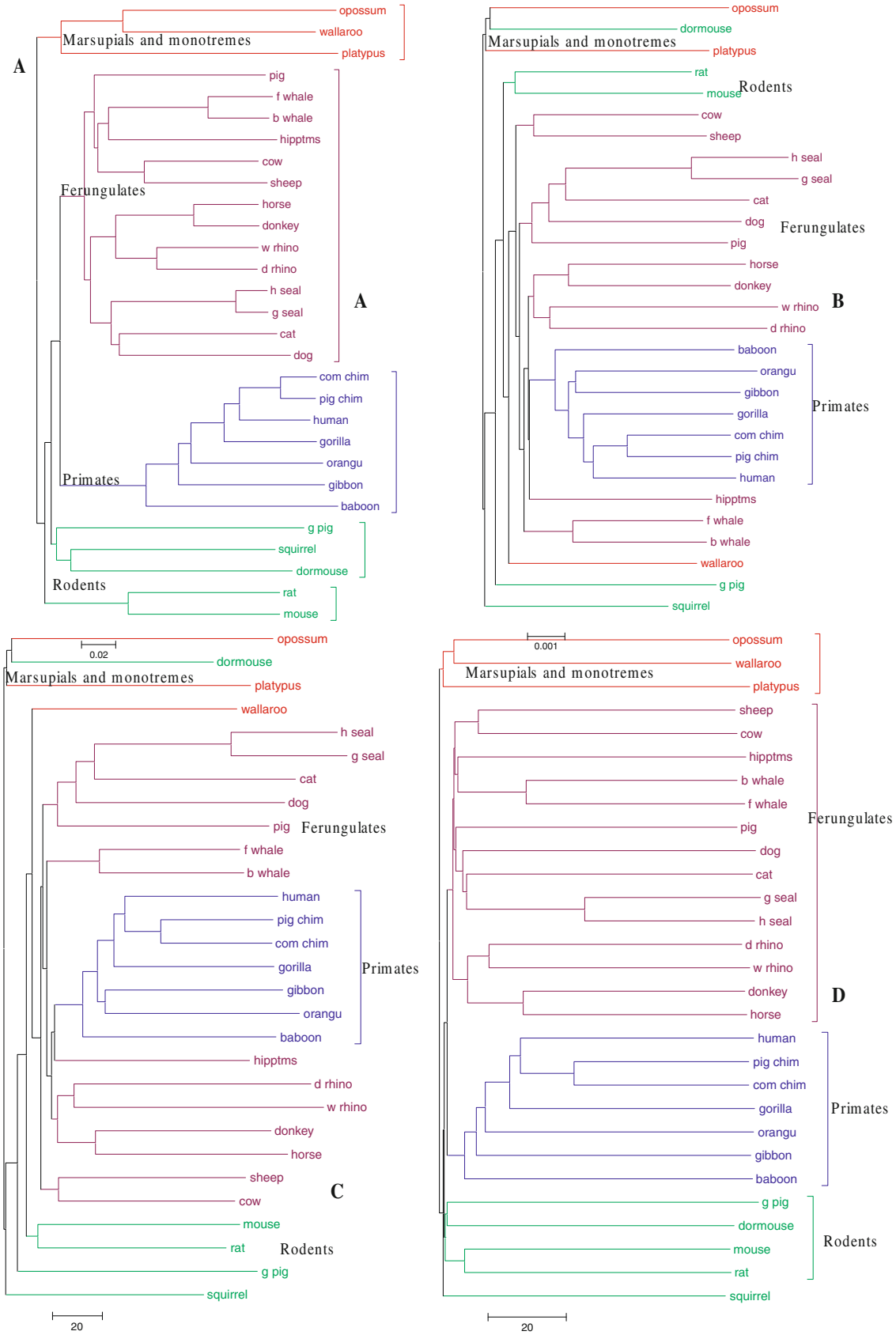
course, this does not happen in practice, and the classes are interspersed. The ROC analysis is used to assess the level of accuracy of this separation without choosing any distance threshold for the separation point. In particular, the AUC will give us a unique number of the relative accuracy of each measure.

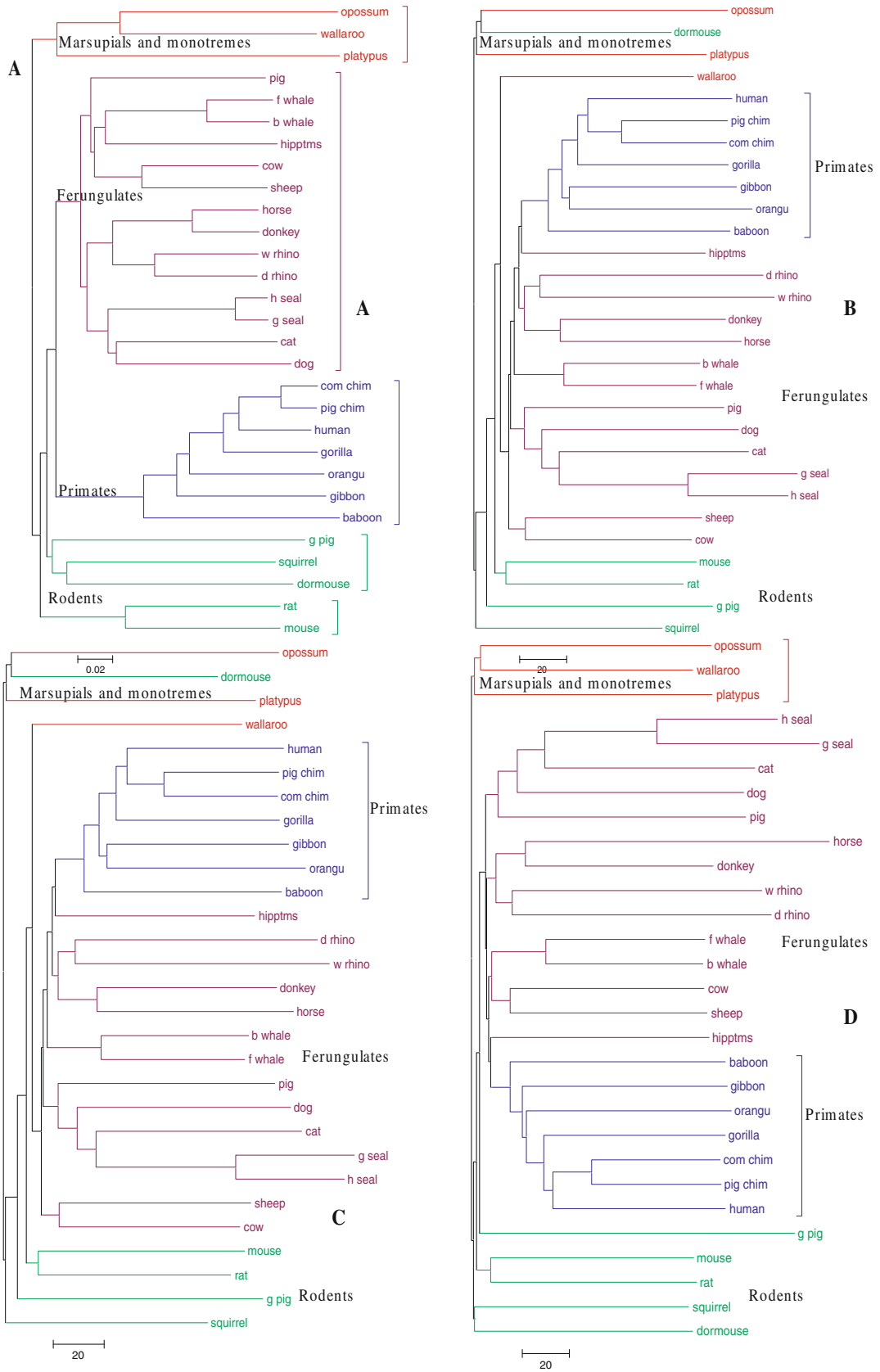
To test the effectiveness of the improved word normalization, we use three standard word-based measures: Euclidean distance (Blaisdell 1986), Cosine distance (Stuart et al. 2002) and standard Euclidean distance (Wu et al. 1997). All standard word-based measures and the standard word-based measures with the improved composition vector (ICV) run with word length *k* from 2 to 7. All standard word-based measures with improved normalization (NF) run with background models of Markov order *r* from 0 to 6 and word length *k* from 2 to 7. For each measures, separate tests are done with all combinations of parameter values, and the best combination is chosen to represent that score in the performance. AUC is computed to evaluate and compare the performances of all the measures. The comparison of AUCs obtained by different measures is presented in Fig. 1.

In the BLASTODERM experiment, the Euclidian distance*ed* with the improved word normalization NF performs better than other measures, with the area under ROC curve 0.7483. The next best measure is the standard Euclidean distance *sd* with the improved word normalization NF. In the PNS experiment, the measure *ed* with the ICV is better than all other measures, its area under ROC curve is 0.7955. In the Fly TRACHEAL experiment, the measure *ed* with the proposed word normalization NF significantly outperforms other methods, and its area under ROC curve is 0.9992, followed by the measure *se* with the proposed word normalization NF. This is a highly significant result, which demonstrates that the measure *ed* with the proposed word normalization NF is successful at detecting the functional similarity of tracheal CRMs. In the EYE experiment, the area under ROC curve of the *ed* with the proposed word normalization NF is 0.7962, significantly better than other statistical measures. In the MUSCLE experiment, the measure *ed* with the proposed word normalization NF significantly outperforms other methods, and its area under ROC curve is 0.8371. It is followed by the *sd* with the proposed word normalization NF. In the LIVER experiments, the measure *sd* with the proposed word normalization NF performs significantly better than other measures. The next best measure is the *ed* with the proposed word normalization NF. In the HBB experiments, the measure *ed* with the proposed word normalization NF achieves the best performance, followed by the *sd* with the proposed word normalization NF.

From Fig. 1, we can see that the word-based measures with the proposed word normalization NF are more efficient, and the proposed word normalization NF seems to take better advantage of the ICV methodology. Although the proposed word normalization NF is similar to ICV, it

Fig. 3 Phylogenetic tree for the complete mammalian mtDNA ▶ sequences of 29 mammalian species, drawn by Treeview program. **a** Obtained by Clustal W, **b** is obtained from standard Euclidean distance, **c** is obtained from standard Euclidean distance with improved composition vector (ICV), **d** is obtained from standard Euclidean distance with the proposed word normalization (NF)


A
opossum
wallaroo
platypus
Marsupials and monotremes
pig
f whale
b whale
hipptms
cow
sheep
Ferungulates
horse
donkey
w rhino
d rhino
h seal
g seal
cat
dog
A
com chim
pig chim
human
gorilla
orangu
gibbon
baboon
Primates
g pig
squirrel
dormouse
Rodents
rat
mouse
B
opossum
dormouse
Marsupials and monotremes
platypus
rat
mouse
Rodents
cow
sheep
h seal
g seal
cat
dog
pig
Ferungulates
horse
donkey
w rhino
d rhino
B
baboon
orangu
gibbon
gorilla
com chim
pig chim
human
Primates
hipptms
f whale
b whale
wallaroo
g pig
squirrel
C
0.02
opossum
dormouse
Marsupials and monotremes
platypus
wallaroo
h seal
g seal
cat
dog
pig
Ferungulates
f whale
b whale
human
pig chim
com chim
gorilla
gibbon
orangu
baboon
Primates
hipptms
d rhino
w rhino
donkey
horse
sheep
cow
mouse
rat
Rodents
g pig
squirrel
20
D
0.001
opossum
wallaroo
Marsupials and monotremes
platypus
sheep
cow
hipptms
b whale
f whale
pig
Ferungulates
dog
cat
g seal
h seal
d rhino
w rhino
donkey
horse
D
human
pig chim
com chim
gorilla
orangu
gibbon
baboon
Primates
g pig
dormouse
mouse
rat
Rodents
squirrel
20


A
opossum
wallaroo
platypus
Marsupials and monotremes
pig
f whale
b whale
hipptms
cow
sheep
Ferungulates
horse
donkey
w rhino
d rhino
h seal
g seal
cat
dog
A
com chim
pig chim
human
gorilla
orangu
gibbon
baboon
Primates
g pig
squirrel
dormouse
Rodents
rat
mouse
B
opossum
dormouse
Marsupials and monotremes
platypus
wallaroo
human
pig chim
com chim
gorilla
gibbon
orangu
baboon
Primates
hipptms
d rhino
w rhino
donkey
horse
b whale
f whale
pig
dog
cat
g seal
h seal
Ferungulates
sheep
cow
mouse
rat
g pig
Rodents
squirrel
C
0.02
opossum
dormouse
Marsupials and monotremes
platypus
wallaroo
human
pig chim
com chim
gorilla
gibbon
orangu
baboon
Primates
hipptms
d rhino
w rhino
donkey
horse
b whale
f whale
Ferungulates
pig
dog
cat
g seal
h seal
sheep
cow
mouse
rat
g pig
Rodents
squirrel
20
D
20
opossum
wallaroo
Marsupials and monotremes
platypus
h seal
g seal
cat
dog
pig
horse
donkey
w rhino
d rhino
f whale
b whale
Ferungulates
cow
sheep
hipptms
baboon
gibbon
orangu
gorilla
com chim
pig chim
human
Primates
g pig
mouse
rat
squirrel
Rodents
dormouse
20

◀ **Fig. 4** Phylogenetic tree for the complete mammalian mtDNA sequences of 29 mammalian species, drawn by Treeview program. **a** Obtained by Clustal W, **b** is obtained from Euclidean distance, **c** is obtained from Euclidean distance with improved composition vector (ICV), **d** is obtained from Euclidean distance with the proposed word normalization (NF)

estimates the expectation and variance of the *k*-word frequencies according to the observed sequences under Markov model rather than assuming the four bases occur randomly with equal chance.

Classification of human exons and introns

The prediction of genes and the classification of coding and noncoding DNA sequences are popular research areas. Numerous advanced statistical algorithms have been developed for gene finding in the past 20 years. They operate on a basic assumption that every exon in a genome should have some distinct sequence features or properties that can distinguish it from the surrounding regions, such as introns or intergenic regions. These algorithms have been reviewed by Fickett and Tung (1992), Fickett (1996) and Guigo (1999). Although good results have been obtained in the recognition of coding and noncoding regions of prokaryotes gene, the statistical features are not sufficient to identify exons in humans because of their limited average length. The classification of the coding and noncoding sequences in humans is still a difficult problem in bioinformatics.

The improved word normalization is further tested to classify human exons and introns. To assess the performance on classifications of the human exons and introns, we construct data sets as follows: 1,200 human exons and 1,200 human introns are extracted from the human exon and intron data (http://bit.uq.edu.au/altExtron/ for human exon and intron datasets), and they are randomly divided into four sets separately. The set of the exons is taken as the 'positive' set, and the set of the introns, is taken as the 'negative' set.

Similar to the above two experiments, the evaluation is based on a binary classification of each sequence pair, where 1 corresponds to the pairs from positive set, 0 corresponds to the pairs from negative set. The ROC analysis are used to assess the effectiveness of alignment-based measures and word-based measures. We change parameter range, because the larger data set is particularly relevant to the problem of computational load. All word-based measures without the normalization and with improved composition vector (ICV) run with the word length *k* from 2 to 6. All word-based measures with the improved normalization (NF) run with background models of Markov order *r* from 0 to 5 and the word length *k* from 2 to 6. The comparison of AUCs for four data sets is presented in Fig. 2.

In terms of the discriminative power of the standard word-based measures with or without word normalization, the measure *ed* with the proposed word normalization NF achieves best performance, with AUC value ranging from 0.9017 to 0.9307 for the four classification tasks. These are excellent values, given that a perfect classification has an AUC score of 1. Take a closer look at Fig. 2, we find that the word-based measures *ed* and *sd* with the improved word normalization NF are more efficient than the ones with the ICV and without the word normalization. However, the measure *cos*, using the proposed word normalization NF and the ICV, is not as good as the measure *cos* without the word normalization.

Phylogenetic analysis

Among three standard word-based measures, standard Euclidean distance and Euclidean distance are distance measures. So we can further test the effectiveness of improved word normalization by phylogenetic analysis. Given a set of DNA, their phylogenetic relationship can be obtained through the following main operations: firstly, we use the Markov model to estimate the word normalization(NF) and calculate their similarity distance by using standard word-based distance; secondly, by arranging all the similarity distance into a matrix, we obtain a pairwise distance matrix; finally, we put the pair-wise distance matrix into the neighbor-joining program in the PHYLIP package (Felsenstein 1989).

The phylogeny of eutherian orders has been unresolved due to conflicting results obtained from comparison of whole mtDNA sequences and individual proteins encoded by mtDNA (Cao et al. 1998). We choose a more controversial data set (29 mammalian species, of which five are rodents that have been widely studied in Reyes et al. (2000), Li et al. (2001), Otu and Sayood (2003). These sequences are described in the supplement material.

The phylogenetic trees constructed by standard Euclidean distance with the proposed word normalization (NF) and Euclidean distance with the proposed word normalization (NF) are presented in Figs. 3d and 4d. Generally, an independent method can be developed to evaluate the accuracy of a phylogenetic tree, or the validity of a phylogenetic tree can be tested by comparing it with authoritative ones. Here, we adopt the former one to test the validity of our phylogenetic tree. We also use Clustal W, Euclidean distance, Euclidean distance with improved composition vector (ICV), standard Euclidean distance, and standard Euclidean distance with improved composition vector (ICV) to construct the phylogenetic tree for the same data.

It has been debated which two of the three main groups of placental mammals are more closely related: primate,

ferungulates, and rodents. This is because by the maximum likelihood method, some protein support the [ferungulates, (primates, rodents)] grouping while other proteins support the [rodents, (ferungulates, primates)] grouping (Cao et al. (1998). The standard Euclidean distance with the proposed word normalization (NF) and Euclidean distance with the proposed word normalization (NF) have reconfirmed the hypothesis of [rodents, (primates, ferungulates)].

Figure 3 shows that our tree is quite consistent with the results of Clustal W and the authoritative ones (Cao et al. 1998; Reyes et al. 2000; Li et al. 2001; Otu and Sayood 2003) in the following three aspects: first of all, marsupials and monotremes are grouped closely, but wallraoo are separated from marsupials and monotremes in the standard Euclidean distance(Fig. 3b) and the standard Euclidean distance with improved composition vector (ICV) (Fig. 3c), which is unreasonable. Because wallraoo belongs to marsupials and monotremes; secondly, five rodents are grouped closely, three non-murid rodents (squirrel, dormouse and guinea pig) are separated from murid rodents (rat and mouse), however dormouse is separated from rodents in the standard Euclidean distance (Fig. 3b) and the standard Euclidean distance with improved composition vector (ICV) (Fig. 3c); finally, primates and ferungulates are grouped in a branch respectively, and they clustered with each other. However, primates and ferungulates are placed into a cluster in the standard Euclidean distance (Fig. 3b) and the standard Euclidean distance with improved composition vector (ICV) (Fig. 3c), which does not consist with the authoritative results reported in Cao et al. (1998), Reyes et al. (2000), Li et al. (2001), Otu and Sayood (2003). These results confirm that the word normalization NF achieves better performance in phylogenetic analysis. The same conclusion can be obtained from Fig. 4. Therefore, the proposed word normalization NF that estimates the expectation and variance of *k*-word frequencies with Markov model are more efficient.

Conclusion

One of the major goals of sequence analysis is to compare biological sequences, which could serve as evidence of structural and functional conservation, as well as of evolutionary relations among the sequences. Despite the prevalence of the alignment-based methods, it is also noteworthy that it is computationally intensive and consequently unpractical for querying large data sets. Therefore, considerable efforts have been made to seek for alternative methods for sequence comparison.

Word-based statistical measures is one of the most well-developed alignment-free methods. This work presented a novel word normalization method to improve biological sequence comparison. In contrast to the improved composition vector (ICV) method based on the uniform model, the proposed word normalization uses Markov model to estimate the expectation and variance of the *k*-word frequencies according to the observed biological sequences. In other words, the proposed word normalization has the ability to adjust the background information for similarity measure using Markov model. To compare the effectiveness of the proposed word normalization, extensive tests were performed, including discrimination between functionally related regulatory sequences and unrelated sequences, intron and exon, and phylogenetic analysis. The results demonstrate that the word-based measures with the proposed word normalization NF are more efficient, and the proposed word normalization NF takes better advantage of the ICV methodology, because it estimates the expectation and variance of *k*-word frequencies based on the observed sequences rather than assuming that four bases occur randomly with equal chance.

Overall, our comparison study highlights effectiveness of the word normalization for biological sequence comparison. Thus, this understanding can then be used to guide development of more powerful measures for sequence comparison with future possible improvement on evolutionary, structure and function study.

Acknowledgments The author thanks all the anonymous referees for their valuable suggestions and support. This work is supported by the National Natural Science Foundation of China (61001214, 61003191), and a research grants (Y2100930, Y6100339) from Zhejiang Provincial Natural Science Foundation of China.

References

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402

Blaisdell BE (1986) Ameasure of the similarity of sets of sequences not requiring sequence alignment. Proc Natl Acad Sci USA 83:5155–5159

Bradley AP (1997) The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognit 30:1145–1159

Cao Y, Janke A, Waddell PJ, Westerman M, Takenaka O, Murata S, Okada N, Paabo S, Hasegawa M (1998) Conflict among individual mitochondrial proteins in resolving the phylogeny of eutherian orders. J Mol Evol 47:307–322

Dai Q, Yang YC, Wang TM (2008) Markov model plus k-word distributions: a synergy that produces novel statistical measures for sequence comparison. Bioinformatics 24:2296–2302

Durbin R, Eddy SR, Krogh A, Mitchison G (1998) Biological sequence analysis. Cambridge University Press, Cambridge

Egan JP (1975) Signal detection theory and ROC-analysis. Academic Press, New York

Felsenstein J (1989) PHYLIP-phylogeny inference package (version 3.2). Cladistics 5:164–166

Felsenstein J (1996) Inferring phylogenies from protein sequences by parsimony, distance and likelihood methods. Methods Enzymol 266:418–427

Fichant G, Gautier C (1987) Statistical method for predicting protein coding regions in nucleic acid sequences. Comput Appl Biosci 3:287–295

Fickett JW, Tung CS (1992) Assessment of protein coding measures. Nucleic Acids Res 20:6641–6450

Fickett JW (1996) Finding genes by computer: the state of the art. Trends Genet 12:316–320

Gallo SM et al (2006) REDfly: a regulatory element database for *Drosophila*. Bioinformatics 22:381–383

Green RE, Brenner SE (2002) Bootstrapping and normalization for enhanced evaluations of pairwise sequence comparison. Proc IEEE 90:1834–1847

Guigo R (1999) In: Genetic databases. Academic Press, New York

Hao B, Qi J (2004) Prokaryote phylogeny without sequence alignment: from avoidance signature to composition distance. J Bioinform Comput Biol 2:1–19

Handl J, Knowles J, Kell DB (2005) Computational cluster validation in post-genomic data analysis. Bioinformatics 21:3201–3212

Huelsenbeck JP, Ronquist F (2001) MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17:754–755

Kantorovitz MR, Robinson GE, Sinha S (2007) A statistical method for alignment-free comparison of regulatory sequences. Bioinformatics 23:i249–i255

Komatsu K, Zhu S, Fushimi H, Qui TK, Cai S, Kadota S (2001) Phylogenetic analysis based on 18S rRNA gene and matK gene sequences of *Panax vietnamensis* and five related species. Plant Med 67:461–465

Kumar S, Tamura K, Nei M (2004) MEGA3: integrated software for molecular evolutionary genetics analysis and sequence alignment. Brief Bioinform 5:150–163

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 17:149–154

Liu Z, Meng J, Sun X (2008) A novel feature-based method for whole genome phylogenetic analysis without alignment: application to HEV genotyping and subtyping. Biochem Biophys Res Commun 368:223–30

Lu GQ, Zhang SP, Fang X (2008) An improved string composition method for sequence comparison. BMC Bioinform 9(Suppl 6):S15

Lu L, Li C, Hagedorn CH (2006) Phylogenetic analysis of global hepatitis E virus sequences: genetic diversity, subtypes and zoonosis. Rev Med Virol 16:5–36

Mitrophanov AY, Borodovsky M (2006) Statistical significance in biological sequence analysis. Brief Bioinform 7:2–24

Mohseni-Zadeh S, Brezellec P, Risler JL (2004) Cluster-C: an algorithm for the large-scale clustering of protein sequences based on the extraction of maximal cliques. Comput Biol Chem 28:211–218

Otu HH, Sayood K (2003) A new sequence distance measure for phylogenetic tree construction. Bioinformatics 19:2122–2130

Pham TD, Zuegg J (2004) A probabilistic measure for alignment-free sequence comparison. Bioinformatics 20:3455–3461

Pham TD (2007) Spectral distortion measures for biological sequence comparisons and database searching. Pattern Recognit 40:516–529

Pipenbacher P, Schliep A, Schneckener S, Schonhuth A, Schomburg D, Schrader R (2002) ProClust: improved clustering of protein sequences with an extended graph-based approach. Bioinformatics 18:S182–S191

Reinert G, Schbath S, Waterman MS (2000) Probabilistic and statistical properties of words: an overview. J Comput Biol 7:1–46

Reyes A, Gissi C, Pesole G, Catzeflis FM, Saccone C (2000) Where do rodents fit? Evidence from the complete mitochondrial genome of *Sciurus vulgaris*. Mol Biol Evol 17:979–983

Rijsbergen CJ (1979) Information retireval. Butterworths, London

Robin S, Daudin JJ (1999) Exact distribution of word occurrences in a random sequence of letters. J Appl Prob 36:179–193

Ronquist F, Huelsenbeck JP (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572–1574

Schbath S (2000) An overview on the distribution of word counts in Markov chains. J Comput Biol 7:193–201

Stajich JE et al (2002) The BioPerl Toolkit: Perl Modules for the life sciences. Genome Res 12:1611–1618

Stuart GW, Moffett K, Baker S (2002) Integrated gene and species phylogenies from unaligned whole genome protein sequences. Bioinformatics 18:100–108

Van Helden J (2004) Metrics for comparing regulatory sequences on the basis of pattern counts. Bioinformatics 20:399–406

Vinga S, Almeida J (2003) Alignment-free sequence comparison: a review. Bioinformatics 19:513–523

Waddell PJ, Kishino H, Ota R (2001) A phylogenetic foundation for comparative mammalian genomics. Genome Inform Ser 12:141–154

Waterman MS (1995) Introduction to computational biology: maps, sequences, and genomes: interdisciplinary statistics. Chapman and Hall, Boca Raton

Wu X, Wan X, Wu G, Xu D, Lin G (2006) Phylogenetic analysis using complete signature information of whole genomes and clustered neighbour-joining method. Int J Bioinform Res Appl 2:219–248

Wu TJ, Burke JP, Davison DB (1997) A measure of DNA sequence dissimilarity based on Mahalanobis distance between frequencies of words. Biometrics 53:1431–1439

Wu TJ, Hsieh YC, Li LA (2001) Statistical measures of DNA dissimilarity under Markov chain models of base composition. Biometrics 57:441–448

Yang L, Chang G, Zhang X, Wang T (2010) Use of the Burrows–Wheeler similarity distribution to the comparison of the proteins. Amino Acids 39(3):887–898

Yao YH, Dai Q, Li C, He PA, Nan XY, Zhang YZ (2008) Analysis of similarity/dissimilarity of protein sequences. Proteins 73(4): 864–871