Using Gaussian Model to Improve Biological Sequence Comparison

QI DAI,1 XIAOQING LIU,2 LIHUA LI,1 YUHUA YAO,2 BIN HAN,1 LEI ZHU1

1 Institute for Biomedical Engineering and Instrumentation, Hangzhou Dianzi University, Hangzhou 310018, People’s Republic of China
2 School of Science, Hangzhou Dianzi University; Hangzhou 310018, People’s Republic of China
3 College of Life Sciences, Zhejiang Sci-Tech University, Hangzhou 310018, People’s Republic of China

Received 11 March 2009; Accepted 14 April 2009
DOI 10.1002/jcc.21322
Published online 28 May 2009 in Wiley InterScience (www.interscience.wiley.com).

Abstract: One of the major tasks in biological 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. Numerous efficient methods have been developed for sequence comparison, but challenges remain. In this article, we proposed a novel method to compare biological sequences based on Gaussian model. Instead of comparing the frequencies of k-words in biological sequences directly, we considered the k-word frequency distribution under Gaussian model which gives the different expression levels of k-words. The proposed method was tested by similarity search, evaluation on functionally related genes, and phylogenetic analysis. The performance of our method was further compared with alignment-based and alignment-free methods. The results demonstrate that Gaussian model provides more information about k-word frequencies and improves the efficiency of sequence comparison.

© 2009 Wiley Periodicals, Inc. J Comput Chem 31: 351–361, 2010

Key words: biological sequence; Gaussian distribution; word frequency; similarity search; phylogenetic analysis

Introduction

The abundance of biomolecular sequence information (generated as a result of the ever-increasing number of large-scale sequencing projects), together with a relatively high cost of “wet lab” experimentation, calls for powerful and efficient computational tools as primary means for high-throughput genomic proteomic investigations. Therefore, computational methods of biological sequence analysis become an indispensable part of the modern scientist’s research arsenal.1 In protein studies, the results of sequence similarity searches in databases help generate reasonable hypotheses concerning structural and functional properties of proteins.2,3 On the DNA level, sequence analysis techniques make it possible to identify genes and functional elements in newly sequenced genomes. Phylogenetic analysis4–9 not only provides us evolutionary relations among the sequences but also provides useful information for pharmaceutical researchers to determine which medicinal species share the same medical qualities. But, these efficient computational methods rely heavily on sequence comparison.

Because of the importance of sequence comparison, numerous methods have been developed.10–34 A typical approach to sequence comparison is based on sequence alignment. Waterman10 and Durbin et al.11 provided comprehensive reviews about this method. The search for optimal solutions using alignment-based method encounters difficulties in: (i) computational load with regard to large databases;12 (ii) choosing the scoring schemes.27

Because of the critical limitations of alignment method, the emergence of research into alignment-free methods is apparent and necessary. Many alignment-free methods have been proposed, but they are still in the early development compared with alignment-based method.14–34 Comparison methods based on k-word frequencies may be the most well-developed alignment-free methods. Reinert et al.26 studied the statistical and probabilistic properties of words in sequences, with emphasis on the deductions of exact distributions and evaluation of its asymptotic approximations. Word-based methods were recently reviewed by Vinga and Almeida.27 Among these word-based methods, each sequence is mapped into an n-dimensional vector according to its k-word frequencies. The similarity score between sequences represented in vector spaces is further defined by Euclidean distance,28 Mahalanobis distance,29 Kullback-Leibler Additional Supporting Information may be found in the online version of this article.

Correspondence to: L. H. Li; e-mail: lilh@hdu.edu.cn
Contract/grant sponsor: National Distinguished Young Research Scientist Award; contract/grant number: 60788101
Contract/grant sponsor: National Natural Science Foundation of China; contract/grant numbers: 60801055, 60801054

© 2009 Wiley Periodicals, Inc.
discrepancy,\textsuperscript{30} Cosine distance\textsuperscript{31} between their corresponding vectors. Recently, several novel word-based methods have been designed for sequence comparison, such as D2z\textsuperscript{(32)}, Gdis.k,\textsuperscript{33} D2 and D3.\textsuperscript{34}

This work presents a novel method for biological sequence comparison based on Gaussian model. Instead of comparing the \(k\)-word frequencies of two sequences directly, we evaluate their \(k\)-word frequencies in a probabilistic framework. Our method was evaluated by extensive tests such as similarity search, evaluation on functionally related genes, and phylogenetic analysis. A comparison of performance between the proposed method and several typical alignment-based or alignment-free methods was taken. The results demonstrate that it is a promising word-method for sequence comparison with potential application in improvement on structure and function prediction.

**Gaussian Model for \(k\)-word Frequencies of Biological Sequences**

**Word Statistics**

There is a large body of literatures on word statistics,\textsuperscript{26} where sequences are interpreted as a succession of symbols and are further analyzed by representing the frequencies of its small segments. 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 \(s\), the count of a \(k\)-word \(w\), denoted by \(c(w)\), is the number of occurrence of \(w\) in the sequence \(s\). The standard approach for counting \(k\)-words in a sequence of length \(m\) is to use a sliding window of length \(k\), shifting the frame one base at a time from position 1 to \(m-k+1\). In this method, \(k\)-words are allowed to overlap in the sequence. In this way, a sequence can be represented by an \(n\)-dimensional vector \(C_s^k\) made up of \(k\)-word counts

\[
C_s^k = (c(w_{k,1}), c(w_{k,2}), \ldots, c(w_{k,n})), \tag{1}
\]

where \(n\) is the total number of all possible \(k\)-words. The frequencies of \(k\)-words, \(F_s^k\), can be calculated by

\[
F_s^k = (f(w_{k,1}), f(w_{k,2}), \ldots, f(w_{k,n})) = \left(\frac{c(w_{k,1})}{m-k+1}, \frac{c(w_{k,2})}{m-k+1}, \ldots, \frac{c(w_{k,n})}{m-k+1}\right). \tag{2}
\]

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

\[
C_s^2 = (c(\text{AA}), c(\text{AG}), c(\text{GG}), c(\text{GA})) = (2, 1, 1, 1),
\]

\[
F_s^2 = (f(\text{AA}), f(\text{AG}), f(\text{GG}), f(\text{GA})) = (0.4, 0.2, 0.2, 0.2).
\]

**Test for Normality of \(k\)-word Frequencies**

In asymptotic cases, Gaussian, Poisson, and compound Poisson approximations have been derived for word counts; the type of approximation depends on the word length and on the method of counting word occurrences. In particular, if \(m\) is the length of the sequence, then, for large \(m\), the distribution of counts of a word can be approximated by the normal distribution; this approximation is good when the length of the word is relatively small compared to the sequence length.\textsuperscript{1}

The simplest method of assessing normality is to look at the frequency distribution histogram. For example, the 3-word frequency histogram of HSLIPAS (Human mRNA for lipoprotein lipase) sequence appears in Figure 1. The most important things to look at are the symmetry and peak of the curve. Figure 1 shows that the distribution of 3-word frequencies of HSLIPAS sequence approximately follows the Gaussian distribution. Visual appraisals can only be used as an indication of the distribution and subsequently better methods must be used.

To test formally for normality we use a Kolmogorov-Smirnov test. Kolmogorov-Smirnov test is a goodness-of-fit test for any statistical distribution. The test relies on the fact that the value of the sample cumulative density function is asymptotically normally distributed. To apply the Kolmogorov-Smirnov test, the main operations are as follows: (1) calculate the cumulative frequency (normalized by the sample size) of the observations as a function of class; (2) calculate the cumulative frequency for a true distribution (most commonly, the Gaussian distribution); (3) find the greatest discrepancy between the observed and expected cumulative frequencies, which is called the “\(D\)-statistic.” The Kolmogorov-Smirnov statistic (\(D\)) is defined as

\[
D = \sup_x |F_n(x) - F(x)|, \tag{3}
\]

where \(F(x)\) is a given cumulative distribution function, and \(F_n(x)\) is an empirical distribution function for \(n\) iid observations \(X_i\), which is defined as

\[
F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I_{X_i \leq x}, \tag{4}
\]
Suppose that we have an i.i.d. data \( f(w_k,1), f(w_k,2), \ldots, f(w_k,n) \) with some unknown distribution \( P \) and we would like to test the hypothesis that \( P \) is equal to a Gaussian distribution \( P_0 \), i.e., decide between the following hypotheses:

\[
H_0 : P = P_0, H_1 : P \neq P_0.
\]

The \( p \)-value obtained by Kolmogorov-Smirnov test tells us whether the data is significantly different from the Gaussian distribution or not. We reject the hypothesis if the test is significant at the 0.05 level. That is to say, if \( p < 0.05 \), we reject \( H_0 \), do not reject \( H_0 \) otherwise. We also take the 3-word frequencies of HSLIP AS sequence for example and perform Kolmogorov-Smirnov test. Since the 3-word frequencies of HSLIP AS sequence is a good model for the 3-word frequencies of HSLIP AS sequence. Then, we perform Lilliefors test that the 3-word frequencies of HSLIP AS sequence comes from the distribution \( N(0.0156, 0.0066) \). At the 0.05 significance level, we accept the normality of 3-word frequencies of HSLIP AS sequence with \( p \)-value 0.19537.

**Gaussian Model for k-word Frequencies**

Many methods for sequence comparison are to fix a short word length \( k \), compute the frequencies of all \( k \)-words in each sequence, and assess the similarity of the two frequency vectors. For example, the dissimilarity score between two sequences \( X \) and \( Y \) are the Euclidian distance \( 28 \) or cosine of the angle \( 31 \) between their \( k \)-word frequency vectors \( F_k^X \) and \( F_k^Y \). Sometimes, these simple methods are not satisfying for sequence comparison, because (i) they treat all word types equally, despite that they have different background, and (ii) it does not take into account the fact that, for a given \( k \)-word, the probability is not a linear function of the number of occurrences. To overcome the problems, the Mahalanobis and standard Euclidean distance, which take into account the data covariance structure, were proposed for sequence comparison.\(^{29}\) In this article, we treat the above two problems by using a probabilistic model of \( k \)-word frequencies.

The Kolmogorov-Smirnov test indicates that the \( k \)-word frequencies of biological sequences can be approximated by Gaussian distribution. This approximation is good when the length of the word is relatively small compared to the sequence length.\(^1\) In what follows we will explore sequence comparison method on the basis of the Gaussian distribution of word frequencies.

The Gaussian distribution, also called the normal distribution, is an important family of continuous probability distributions, applicable in many fields. Each member of the family may be defined by two parameters, location and scale: the mean (“average,” \( \mu \)) and variance (standard deviation squared, \( \sigma^2 \)) respectively. The standard Gaussian distribution is the Gaussian distribution with a mean of zero and a variance of one. To indicate that a real-valued random variable \( X \) is normally distributed with mean \( \mu \) and variance \( \sigma^2 \geq 0 \), we write

\[
X \sim N(\mu, \sigma^2).
\]

There are various ways to characterize a probability distribution. The most widely used one is probability density function (PDF). The probability density function of the Gaussian distribution is

\[
\varphi_{\mu, \sigma^2}(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} = \frac{1}{\sigma} \varphi \left( \frac{x - \mu}{\sigma} \right), \quad x \in \mathbb{R},
\]

where \( \sigma > 0 \) is the standard deviation, the real parameter \( \mu \) is the expected value, and

\[
\varphi(x) = \varphi_{0,1}(x) = \frac{e^{-x^2/2}}{\sqrt{2\pi}}, \quad x \in \mathbb{R},
\]
is the density function of the “standard” Gaussian distribution: i.e., the Gaussian distribution with $\mu = 0$ and $\sigma = 1$. The distribution function of the Gaussian distribution is expressed in terms of the density function as follows:

$$
\Phi(\sigma^2;x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(y-\mu)^2}{2\sigma^2}} \, dy = \frac{1}{\sigma \sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(y-\mu)^2}{2\sigma^2}} \, dy \tag{9}
$$

The standard Gaussian distribution function is just the general distribution function evaluated with $\mu = 0$ and $\sigma = 1$:

$$
\Phi(x) = \Phi_{0,1}(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{y^2}{2}} \, dy. \tag{10}
$$

A biological sequence $s$, of length $m$, is defined as a linear succession of symbols from a finite alphabet $\mathcal{A}$, with size of $|\mathcal{A}|$. All possible sequences of length $k$ with symbols from the alphabet $\mathcal{A}$ compose a $k$-word set, which corresponds to a $k$-word frequencies set $\mathcal{F}_k$. Suppose $\mathcal{F}_k$ is a sample space and the frequency of each $k$-word is a random variable denoted by $f_{w_k}$, the frequency of $k$-word is approximately followed by the Gaussian distribution with mean $\mu$ and variance $\sigma^2$. That is to say,

$$
f_{w_k} \sim N(\mu, \sigma^2). \tag{11}
$$

Give two biological sequences $X$ and $Y$, the frequencies of $k$-word $f_{w_k}$ in sequences $X$ and $Y$ follow two different Gaussian models

$$
\begin{align*}
&f_{w_k}^X \sim N(\mu_X, \sigma^2_X), \\
&f_{w_k}^Y \sim N(\mu_Y, \sigma^2_Y).
\end{align*} \tag{12a, 12b}
$$

According to distribution function of the Gaussian distribution, we have

$$
\Phi_{\mu_X,\sigma^2_X}(f_{w_k}^X) = P(x \leq f_{w_k}^X) = \frac{1}{\sigma_X \sqrt{2\pi}} \int_{-\infty}^{f_{w_k}^X} e^{-\frac{(y-\mu_X)^2}{2\sigma^2}} \, dy, \tag{13a}
$$

$$
\Phi_{\mu_Y,\sigma^2_Y}(f_{w_k}^Y) = P(x \leq f_{w_k}^Y) = \frac{1}{\sigma_Y \sqrt{2\pi}} \int_{-\infty}^{f_{w_k}^Y} e^{-\frac{(y-\mu_Y)^2}{2\sigma^2}} \, dy. \tag{13b}
$$

where $f_{w_k}^X$ ($f_{w_k}^Y$) is the frequency of $k$-word $w_k$ in sequence $X$ ($Y$). $\Phi_{\mu_X,\sigma^2_X}(f_{w_k}^X)$ ($\Phi_{\mu_Y,\sigma^2_Y}(f_{w_k}^Y)$) is the probability of observing frequencies of $w_k$ ($\leq f_{w_k}^X$, $\leq f_{w_k}^Y$) in sequence $X$ ($Y$). 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 probability $\Phi_{\mu_X,\sigma^2_X}(f_{w_k}^X)$ ($\Phi_{\mu_Y,\sigma^2_Y}(f_{w_k}^Y)$) measures a level of expression—low value of $\Phi_{\mu_X,\sigma^2_X}(f_{w_k}^X)$ ($\Phi_{\mu_Y,\sigma^2_Y}(f_{w_k}^Y)$) corresponds to low expression of word $w_k$, and large value of $\Phi_{\mu_X,\sigma^2_X}(f_{w_k}^X)$ ($\Phi_{\mu_Y,\sigma^2_Y}(f_{w_k}^Y)$) corresponds to high expression of the word $w_{kj}$ in sequence $X$ ($Y$). We define the probability distance between $X$ and $Y$ as

$$
d_{\text{Dist}}(X, Y) = \sum_{i=1}^{\mathcal{F}} \left| \Phi_{\mu_X,\sigma^2_X}(f_{w_k}^X) - \Phi_{\mu_Y,\sigma^2_Y}(f_{w_k}^Y) \right|
$$

$$
= \sum_{i=1}^{\mathcal{F}} \left| \frac{1}{\sigma_X \sqrt{2\pi}} \int_{-\infty}^{f_{w_k}^X} e^{-\frac{(y-\mu_X)^2}{2\sigma^2}} \, dy - \frac{1}{\sigma_Y \sqrt{2\pi}} \int_{-\infty}^{f_{w_k}^Y} e^{-\frac{(y-\mu_Y)^2}{2\sigma^2}} \, dy \right|
$$

$$
= \sum_{i=1}^{\mathcal{F}} \left| \Phi \left( \frac{f_{w_k}^X - \mu_X}{\sigma_X} \right) - \Phi \left( \frac{f_{w_k}^Y - \mu_Y}{\sigma_Y} \right) \right|. \tag{14}
$$

The $d_{\text{Dist}}(X, Y)$ has the following properties: (i) it is a distance measure, because it satisfies positivity, symmetry and triangle inequality; (ii) background information is incorporated into the measure; (iii) $k$-words with identical frequency in two sequences may have different expression levels.

**Estimate of Parameters**

Since the mean $\mu$ and variance $\sigma^2$ are priori unknown, we have to estimate them according to the observed sequences. Here, we estimate the parameters of Gaussian model by using the maximum likelihood method.

Give a biological sequence, its $k$-word frequencies are $f(w_{k1}), f(w_{k2}), \ldots, f(w_{kn})$. Suppose $f(w_{k1}), f(w_{k2}), \ldots, f(w_{kn})$ are independent and each is normally distributed with expectation $\mu$ and variance $\sigma^2 \geq 0$. These observed values of these $n$ random variables make up a “sample of size $n$ from a normally distributed population.” It is desired to estimate the “population mean” $\mu$ and the “population standard deviation” $\sigma$, based on the observed values of this sample. The continuous joint probability density function of these $n$ independent random variables is

$$
f(x_1, x_2, \ldots, x_n, \mu, \sigma) = \prod_{i=1}^{n} \varphi_{\mu,\sigma^2}(x_i) = \prod_{i=1}^{n} \left\{ \frac{1}{\sigma \sqrt{2\pi}} \int_{-\infty}^{x_i} e^{-\frac{(y-\mu)^2}{2\sigma^2}} \, dy \right\} = \prod_{i=1}^{n} \left\{ \frac{1}{\sigma \sqrt{2\pi}} \int_{-\infty}^{x_i} e^{-\frac{(y-\mu)^2}{2\sigma^2}} \, dy \right\}
$$

As a function of $\mu$ and $\sigma$, the likelihood function based on the observations $f(w_{k1}), f(w_{k2}), \ldots, f(w_{kn})$ is

$$
L(\mu, \sigma) = C e^{-\frac{\sum_{i=1}^{n} \left( f_{w_k}^X - \mu \right)^2}{2\sigma^2}}, \mu, \sigma > 0, \tag{16}
$$

with some constant $C > 0$.

In the method of maximum likelihood, the values of $\mu$ and $\sigma$ that maximize the likelihood function are taken as estimates of the population parameters $\mu$ and $\sigma$. Usually in maximizing a function of two variables, one might consider partial derivatives. But here we will exploit the fact that the value of $\mu$ that maximizes the likelihood function with $\sigma$ fixed does not depend on $\sigma$. Therefore, we can find
that value of $\mu$, then substitute it for $\mu$ in the likelihood function, and finally find the value of $\sigma$ that maximizes the resulting expression.

It is evident that the likelihood function is a decreasing function of the sum

$$\sum_{i=1}^{n} (f(w_{ki}) - \mu)^2.$$  \hfill (17)

So we want the value of $\mu$ that minimizes this sum. Let

$$\tilde{f} = (f(w_{k1}) + \cdots + f(w_{kn}))/n$$  \hfill (18)

be the “sample mean” based on the $n$ observations. Observe that

$$\sum_{i=1}^{n} (f(w_{ki}) - \mu)^2 = \sum_{i=1}^{n} (f(w_{ki}) - \tilde{f} + \tilde{f} - \mu)^2$$

$$= \sum_{i=1}^{n} (f(w_{ki}) - \tilde{f})^2 + 2\tilde{f}(\mu - \tilde{f})\sum_{i=1}^{n} (f(w_{ki}) - \tilde{f}) + \sum_{i=1}^{n} (\tilde{f} - \mu)^2$$

$$= \sum_{i=1}^{n} (f(w_{ki}) - \tilde{f})^2 + n(\tilde{f} - \mu)^2.$$  \hfill (19)

Only the last term depends on $\mu$ and it is minimized by

$$\hat{\mu} = \tilde{f}.$$  \hfill (20)

That is the maximum-likelihood estimate of $\mu$ based on the $n$ observations $f(w_{k1}), f(w_{k2}), \ldots, f(w_{kn})$. When we substitute that estimate for $\mu$ into the likelihood function, we get

$$L(\tilde{f}, \sigma) = \frac{C}{\sigma^n} e^{-\left(\frac{\sum_{i=1}^{n} (f(w_{ki}) - \tilde{f})^2}{2\sigma^2}\right)}, \quad \sigma > 0,$$  \hfill (21)

It is conventional to denote the “log-likelihood function,” i.e., the logarithm of the likelihood function, by a lower-case $\ell$, and we have

$$\ell(\tilde{f}, \sigma) = \log C - n \log \sigma^n - \frac{\sum_{i=1}^{n} (f(w_{ki}) - \tilde{f})^2}{2\sigma^2}, \quad \sigma > 0,$$  \hfill (22)

and then

$$\frac{d\ell(\tilde{f}, \sigma)}{d\sigma} = \frac{-n}{\sigma} + \frac{\sum_{i=1}^{n} (f(w_{ki}) - \tilde{f})^2}{\sigma^3}$$

$$= \frac{n}{\sigma^3} \left(\sigma^2 - \frac{1}{n} \sum_{i=1}^{n} (f(w_{ki}) - \tilde{f})^2\right), \quad \sigma > 0.$$  \hfill (23)

Thus

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (f(w_{ki}) - \tilde{f})^2.$$  \hfill (24)

\section*{Evaluation}

The proposed method is evaluated by extensive experiments such as similarity search, evaluation on functionally related genes, and phylogenetic analysis. We presently grouped our experiments into two sets. The first one, performed via ROC (receiver operating curve) analysis, aims at assessing the intrinsic ability of our method to search for similar sequences from a database and discriminate functionally related genes from unrelated sequences. The second one aims at assessing how well our method is used for phylogenetic analysis.

\subsection*{Evaluation Method}

The method that will be used here to evaluate performance of the present method is based on the analysis of ROC curves. ROC goes back to signal detection and classification problems and is now widely used.\textsuperscript{37} 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), also named sensitivity or coverage versus false positives (FP), or (1-specificity), encountered for each threshold, properly normalized [eq. (25)].

$$\text{sensitivity} = \frac{\text{True Positives}}{\text{Positives}} = \frac{TP}{TP + FN},$$

$$\text{specificity} = \frac{\text{True Negatives}}{\text{Negatives}} = \frac{TN}{TN + FP},$$

$$1 - \text{specificity} = \frac{FP}{TN + FP}.$$  \hfill (25)

A ROC curve is simply the plot of sensitivity versus (1-specificity) for different threshold values. The area under a ROC curve (AUC) is a widely employed parameter to quantify the quality of a classifier because it is a threshold independent performance measure and is closely related to the Wilcoxon signed-rank test.\textsuperscript{38} For a perfect classifier, the AUC is 1 and for a random classifier the AUC is 0.5. For additional results and comprehensive discussion on AUC measure, see ref. 39.

\subsection*{Similarity Search}

The proposed method is used to search for similar sequences of a query sequence from a database of 39 library sequences, of which 20 sequences are known to be similar in biological function to the query sequence, and the remaining 19 sequences are known as being not similar in biological function to the query sequence. This data set has been studied in refs. 12, 30 and 40. These 39 sequences were selected from mammals, viruses, plants, etc., of which lengths vary from 322 to 14121 bases. The query sequence is HSLIPAS (Human mRNA
Figure 3. ROC curves for similarity search dataset. Similarity measure names are presented with word length as suffix. A random classifier would generate equal proportions of FP and TP classifications, which corresponds to the ROC diagonal (dashed line).

The 20 sequences, which are known as being similar in biological function to HSLIP AS, are as follows: OOLPLIP (Oestrus ovis mRNA for lipoprotein lipase, 1656 bp), SSLPLRNA (pig back fat Sus scrofa cDNA similar to S. scrofa LPL mRNA for lipoprotein lipase, 2963 bp), RATLLIPA (Rattus norvegicus lipoprotein lipase mRNA, complete cds, 3617 bp), MUSLIPLIP (Mus musculus lipoprotein lipase gene, partial cds, 3806 bp), GPLLPL (guinea pig lipoprotein lipase mRNA, complete cds, 1744 bp), GGLPL (chicken mRNA for adipose lipoprotein lipase, 2328 bp), HSHTGL (human mRNA for hepatic triglyceride lipase, 1603 bp), HUMLIPH (human hepatic lipase mRNA, complete cds, 1550 bp), HUMLIPH06 (human hepatic lipase gene, exon 6, 322 bp), RATHLP (rat hepatic lipase mRNA, 1639 bp), RABTRIL [Oryctolagus cuniculus (clone TGL-5K) triglyceride lipase mRNA, complete cds, 1444 bp], ECPL (Equus caballus mRNA for pancreatic lipase, 1443 bp), DOGPLIP (canine lipase mRNA, complete cds, 1493 bp), DMYOLK [Drosophila gene for yolk protein I (vitellogenin), 1723 bp], BOVLDLR [bovine low-density lipoprotein (LDL) receptor mRNA, 879 bp], HSBMHSP (Homo sapiens mRNA for basement membrane heparan sulfate proteoglycan, 13,790 bp), HUMAPOAICI (human apolipoprotein A-I and C-III genes, complete cds, 8966 bp), RABVLDLR (O. cuniculus mRNA for very LDL receptor, complete cds, 3209 bp), HSLDL100 (human mRNA for apolipoprotein B-100, 14,121 bp), and HUMAPOBF (human apolipoprotein B-100 mRNA, complete cds, 10,089 bp).

The other 19 sequences known as being not similar in biological function to HSLIP AS are as follows: A1MVRNA2 [alfalfa mosaic virus (A1M4) RNA 2, 2593 bp], AAHAV33A [Acanthocheilonema vitaeae pepsin-inhibitorlike-protein (Av33) mRNA sequence, 1048 bp], AA2CG (adenovirus-associated virus 2, complete genome, 4675 bp), ACVPBD64 (artificial cloning vector plasmid BD64, 4780 bp), AL3HP (bacteriophage alpha-3 H protein gene, complete cds, 1786 bp), AABADA [Aedes aegypti abd-A gene for abdominal-A protein homolog (partial), 1759 bp], BACBDGALA [Bacillus circulans betad-galactosidase (bgaA) gene, complete cds, 2555 bp], BBCA (Bos taurus mRNA for cyclin A, 1512 bp), BCP1 (bacteriophage Cbp1 genome DNA, complete sequence, 4877 bp) and CHIBATPB (sweet potato chloroplast F1-ATPase beta and epsilon-subunit genes, 2007 bp), A7NIFH (Anabaena 7120 nifH gene, complete CDS, 1271 bp), AA16S (Amycolatopsis azurea 16S rRNA, 1300 bp), ABGACT2 (Absidia glauca actin mRNA, complete cds, 1309 bp), ACTIBETLC (Actinomadura R39 DNA for beta-lactamase gene, 1902 bp), AMTG65RNRA (Ambystoma mexicanum AxU1 snRNA gene, complete sequence, 1027 bp), ARST18B (cloning vector pAST 18b for Caenorhabditis elegans, 3052 bp), GCALIP2 (Geotrichum candidum mRNA for lipase II precursor, partial cds, 1767 bp), ACGGGLINE (Ateles geoffroyi gamma-globin gene and L1 LINE element, 7360 bp), and HUMCAN (H. sapiens CaN19 mRNA sequence, 427 bp).

ROC curves are computed to evaluate and compare the performance of our measure with other measures. The evaluated measures are as follows: the similarity measures based on Clustal W, Euclidean distance (eu), Mahalanobis distance (md), standard Euclidean distance (sd), Kullback-Leibler discrepancy (kld), Cosine distance (cos), D2z, D2, D3 and our measure dNor. All measures based on k-word frequencies run with k from 2 to 5. For each measures, separate tests are done with each combination of parameter values, and the best combination is chosen to represent that score in the performance. The ROC curves obtained for the similarity search are presented in Figure 3.

The AUC value is typically used as a measure of overall discrimination accuracy. Table 1 provides the areas under ROC curves...
Table 1. Comparison of AUCs Obtained from All the Similarity Measures with Word Length as Suffix.

| Methods    | Area    | TStd Error | Asymptotic | Lower bound | Upper bound |
|------------|---------|------------|------------|-------------|-------------|
| Clustal W  | 0.922   | 0.048      | 0.000      | 0.827       | 1.017       |
| cos.2      | 0.900   | 0.053      | 0.000      | 0.796       | 1.004       |
| ed.3       | 0.897   | 0.058      | 0.000      | 0.785       | 1.010       |
| kld.4      | 0.900   | 0.051      | 0.000      | 0.800       | 1.000       |
| D2z.5      | 0.929   | 0.040      | 0.000      | 0.851       | 1.007       |
| md.3       | 0.916   | 0.050      | 0.000      | 0.819       | 1.013       |
| sd.4       | 0.705   | 0.089      | 0.028      | 0.531       | 0.879       |
| D2,4       | 0.874   | 0.058      | 0.000      | 0.760       | 0.987       |
| D3,3       | 0.889   | 0.054      | 0.000      | 0.783       | 0.996       |
| dNor.4     | 0.953   | 0.030      | 0.000      | 0.894       | 1.011       |

(AUC) obtained from all the measures. Figure 3 and Table 1 show that dNor.4 performs better than other alignment-based or alignment-free measures on similarity search. Its area under ROC curve is 0.953 with the small standard error 0.030 for this estimate. Clustal W outperforms other alignment-free measures such as cos.3, ed.3, kld.4, md.3, sd.4, D2,4, and D3,3. Among the similarity measures based on k-word distributions, D2z.5 is clearly more efficient than other measures. The main surprise of this analysis is that when we explore the distribution information of k-word frequencies in our way, dNor.4 performs better than other similarity measures based on k-word frequencies. The inspection of the ROC curves themselves (Fig. 3) further illustrates this comparison between similarity measures.

Evaluation on Functionally Related Genes

The proposed Gaussian model of k-word frequencies is further tested to evaluate if functionally or evolutionarily related gene pairs are scored better than unrelated pairs of random sequences. To assess the performance on functionally related genes, we construct data sets as follows. We selected three sets of genes, each involved in a particular pathway: nitrogen metabolism (NIT family, 31 genes), phosphate utilization (PHO family, 13 genes), and methionine biosynthesis (MET family, 20 genes). They are well studied in ref. 41. We retrieved the 800 bp sequence upstream the start codon of each gene as “positive” sets. As “negative” sets, we generated random sequences with lengths matching the sequence in “positive” sets.

Each pair of sequences in the positive set is compared, and so is each pair in the negative set. The evaluation procedure 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(n choose 2), which is used as prediction. Also, we can get a vector of length 2(n choose 2) consisting of 1 and 0 as class labels. A perfect measure would completely separate negative from positive set. Of course, this does not happen in practice, and the classes are interspersed. The ROC curves permit 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.

The similarity measures evaluated here are as follows: the similarity measures based on alignment, Euclidean distance (eu), Mahalanobis distance (md), standard Euclidean distance(sd), Kullback Leibler discrepancy(kld), Cosine distance (cos), D2z, D2, D3, and our measure dNor, where the similarity measures based on alignment are Needleman-Wunsch (global alignment) or Smith-Waterman (local alignment) raw scores, with no correction for statistical significance, using linear gap penalties or affine gap penalties, with a gap penalty of 2. All measures based on k-word distributions run with k from 2 to 5. For each measures, separate tests are done with each combination of parameter values, and the best combination is chosen to represent that score in the performance. ROC curves are computed to evaluate and compare the performances of our measures and other measures. The ROC curves obtained for NIT, Met, and PHO are presented in Figures 4 and 5.

Table 2 summarizes the AUCs obtained from all the measures for three data sets. In the MET experiment, dNor.4 performs better than other alignment-based and word-based measures, with the area under ROC curve 1.000. The next best measure is kld.2, and the other measures lag behind. In the NIT experiment, dNor.5 measure is better than all other measures, and its area under ROC curve is 1.000. In the MUSCLE experiment, dNor.3 outperforms other methods, with the area under ROC curve 1.000. It is followed by kld.5. From the three experiments, we can see that dNor, exploring the distribution of k-word frequencies, performs better than other measures. The inspection of the ROC curves themselves (Figs. 4 and 5) further illustrates these comparisons among similarity measures. The highly significant results of our method demonstrate that the dNor measure is successful at detecting the functional similarity of genes from the random sequences. Since the different genes are only functionally related and not orthologous, the gene search algorithm requires a method that can discern functional similarity among candidate genes based on their sequence similarity. From Table 2, we can note that the alignment-based methods lag behind some alignment-free methods.

Construction of Phylogenetic Tree of Coronavirus

Since the outbreak of atypical pneumonia referred to as severe acute respiratory syndrome (SARS), more attentions have been paid to the relationships between the SARS-CoVs and the other coronaviruses, which would be helpful to discover drugs and develop vaccines against the virus. Generally, coronaviruses can be divided into three groups according to serotypes. Group I and group II contain mammalian viruses, while group II coronaviruses contain a hemagglutinin esterase gene homologous to that of Influenza C virus. Group III contains only avian.

Based on the Gaussian model of k-word frequencies, we next consider to infer the phylogenetic relationships of coronaviruses with the complete coronavirus genomes. The 24 complete coronavirus genomes used in this article were downloaded from GenBank, of which 12 are SARS-CoVs and 12 are from other groups of coronaviruses. The name, accession number, abbreviation, and genome length for the 24 genomes are listed in Table 3. Given a set...
Figure 4. ROC curves for data sets Met and NIT. ROC(a) curves for data MET and ROC(b) curves for data NIT. Similarity measure names are presented with word length as suffix. A random classifier would generate equal proportions of FP and TP classifications, which corresponds to the ROC diagonal (dashed line).

Figure 5. ROC curves for data sets PHO. Similarity measure names are presented with word length as suffix. A random classifier would generate equal proportions of FP and TP classifications, which corresponds to the ROC diagonal (dashed line).
of biological sequences, their phylogenetic tree can be obtained through the following main operations: firstly, we construct the Gaussian model for biological sequences; secondly, we calculate their similarity degree by using our measure $d_{Nor}$. Thirdly, by arranging all the similarity degree into a matrix, we obtain a pair-wise distance matrix. Finally, we put the pair-wise distance matrix into the neighbor-joining program in the PHYLIP package. We obtain the phylogenetic relationships drawn by MEGA program. In Figure 6, we present the unrooted phylogenetic tree belonging to 24 species.

Figure 6 shows that our results are quite consistent with the accepted taxonomy and authoritative ones in the following four aspects. First, all SARS-CoVs are grouped in a separate branch, which appear different from the other three groups of coronaviruses. Secondly, BCOV, BCOVL, BCOVM, BCOVQ, MHV, MHV2, MHVM, and MHVP are grouped into a branch, which is consonant with that they belong to group II. Thirdly, HCoV-229E, TGEV, and PEDV are closely related to each other, which is consistent with the fact that they belong to group I. Finally, IBV forms a distinct branch within the genus Coronavirus, because it belongs to group III. Grigoriev found that the mutational patterns in SARS-CoV genome were strikingly different from the other coronaviruses in terms of mutation rates. Phylogenetic analysis based on codon usage pattern suggested that SARS-CoV was diverged far from all the three known groups of coronavirus. Rota et al. found out that the overall level of similarity between SARS-CoVs and the other coronaviruses is low. Our tree also reconfirms that SARS-CoVs are not closely related to any previously isolated coronaviruses and form a new group, which indicates that the SARS-CoVs have undergone an independent evolution path after the divergence from the other coronaviruses.

Whole genome-based phylogenetic analysis is appealing because single gene sequences generally do not possess enough information to construct an evolutionary history of organisms. Now phylogenetic analysis based on sequence alignments is well developed. However, it can hardly be applied to complete genomes, because the computational load of multiple alignment increases with the increasing length of sequence. Being different from the sequence alignment method, the current method is more simple and yields results reasonably.

### Table 2. Comparison of AUCs Obtained from All the Similarity Measures with Word Length as Suffix.

| Method | MET | NIT | PHI |
|--------|-----|-----|-----|
| NW-linear | 0.549 | 0.531 | 0.568 |
| NW-affine | 0.551 | 0.545 | 0.586 |
| SW-linear | 0.823 | 0.805 | 0.868 |
| SW-affine | 0.823 | 0.805 | 0.868 |
| cos.2 | 0.817 | 0.816 | 0.888 |
| eu.2 | 0.881 | 0.879 | 0.849 |
| kld.2 | 0.991 | 0.988 | 0.972 |
| D2z.2 | 0.812 | 0.854 | 0.847 |
| md.3 | 0.959 | 0.922 | 0.902 |
| sd.3 | 0.954 | 0.915 | 0.896 |
| D2.2 | 0.838 | 0.785 | 0.849 |
| dNor | 1.000 | 1.000 | 1.000 |

### Table 3. The Accession Number, Abbreviation, Name, and Length for Each of the 24 Coronavirus Genomes.

| No. | Accession | Group | Abbreviation | Genome | Length (nt) |
|-----|-----------|-------|--------------|--------|-------------|
| 1   | NC_002645 | I     | HCoV-229E    | Human coronavirus 229E | 27,317 |
| 2   | NC_002306 | I     | TGEV         | Transmissible gastroenteritis virus | 28,586 |
| 3   | NC_003436 | I     | PEDV         | Porcine epidemic diarrhea virus | 28,033 |
| 4   | U00735    | II    | BCOV         | Bovine coronavirus strain Mebus | 31,032 |
| 5   | AF391542  | II    | BCOVL        | Bovine coronavirus isolate BCoV-LUN | 31,028 |
| 6   | AF220295  | II    | BCOVQ        | Bovine coronavirus strain Quebec | 31,100 |
| 7   | NC_003045 | II    | BC0V         | Bovine coronavirus | 31,026 |
| 8   | AF208067  | II    | MHVM         | Murine hepatitis virus strain ML-10 | 31,233 |
| 9   | AF201929  | II    | MHV2         | Murine hepatitis virus strain 2 | 31,276 |
| 10  | AF208066  | II    | MHVP         | Murine hepatitis virus strain Penn 97-1 | 31,112 |
| 11  | NC_001846 | II    | MHV          | Murine hepatitis virus | 31,357 |
| 12  | NC_001451 | III   | IBV          | Avian infectious bronchitis virus | 27,608 |
| 13  | AY728488  | –     | BJ01         | SARS coronavirus BJ01 | 29,725 |
| 14  | AY278741  | –     | Urbani       | SARS coronavirus Urbani | 29,727 |
| 15  | AY278491  | –     | HKU-39849    | SARS coronavirus HKU-39849 | 29,742 |
| 16  | AY278554  | –     | CUHK-W1      | SARS coronavirus CUHK-W1 | 29,736 |
| 17  | AY282752  | –     | CUHK-Su10    | SARS coronavirus CUHK-Su10 | 29,736 |
| 18  | AY283794  | –     | SIN2500      | SARS coronavirus Sin2500 | 29,711 |
| 19  | AY283795  | –     | SIN2677      | SARS coronavirus Sin2677 | 29,705 |
| 20  | AY283796  | –     | SIN2679      | SARS coronavirus Sin2679 | 29,711 |
| 21  | AY283797  | –     | SIN2748      | SARS coronavirus Sin2748 | 29,705 |
| 22  | AY283798  | –     | SIN2774      | SARS coronavirus Sin2774 | 29,711 |
| 23  | AY291451  | –     | TW1          | SARS coronavirus TW1 | 29,729 |
| 24  | NC_004718 | –     | TOR2         | SARS coronavirus | 29,751 |
IBV

BCOV5

BCOV5

VOC

MHV5

MHV

MHV2

MHV5p

CUHK-W1

CUHK-Su10

BJ01

HKU-39849

Urban

TW1

SIN2679

SIN2677

SIN2500

SIN2748

SIN2774

PEDV

TGEV

HCoV-229E

Figure 6. The unrooted consensus species tree for 24 coronavirus by our distance \( d_{\text{mfs}} \) at \( k = 6 \) using whole genomes.

Conclusion

Sequence comparison is rapidly becoming an essential tool for bioinformatics applications. It has been used to support other types of analyses, from searching a database with a query DNA sequence to the phylogenetic tree construction. Despite the prevalence of alignment-based methods, it is noteworthy that alignment-based method is computationally intensive and consequently unpractical for querying large data sets, which forces the use of some heuristics to reduce the running times, as exemplified by BLAST. Alignment-free comparison method is therefore of great value as it reduces the technical constraints of alignments.

A novel alignment-free method for sequence comparison is proposed in this work. We assume that the frequencies of a given \( k \)-word in a biological sequence follows the Gaussian distribution. The similarity between two sequences can be evaluated by the difference between their corresponding Gaussian models. In contrast to the traditional word-based methods based on frequencies of fixed \( k \)-words, our method takes distribution information of \( k \)-word frequencies into account. In other words, our method has the ability to adjust the background information for similarity measure using \( k \)-word frequencies. The test of our methods are to perform similarity search and evaluate the functionally related genes. To evaluate this method, we compare it with alignment-based or word-based methods. The results also indicated that it is a necessity for alignment-free methods to extract more information in order to have a good comparison performance. This understanding can then be used to guide development of more powerful sequence comparison method for potential improvement on evolutionary study, structure and function prediction.

Acknowledgment

The authors thank all the anonymous referees for their valuable suggestions and support.

References

1. Mitrophanov, A. Y.; Borodovsky, M. Brief Bioinform 2006, 7, 2.
2. Altschul, S. F.; Madden, T. L.; Schaffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J. Nucleic Acids Res 1997, 25, 3389.
3. Felsenstein, J. Methods Enzymol 1996, 266, 418.
4. Waddell, P. J.; Kishino, H.; Ota, R. Genome Inform Ser 2001, 12, 141.
5. Randić, M.; Guo, X.; Balaban, A. T.; Czekanowski, L. In Chemical Graph Theory; Chapman & Hall: New York, 1995.
6. Vukicevic, D. J Comput Chem 2008, 29, 1632.
7. Vukicevic, D.; Dai, Q.; Nan, X. Y.; He, P. A.; Nie, Z. M.; Zhou, S. P.; Zhang, Y. Z. J Comput Chem 2008, 29, 1632.
8. Dai, Q.; Wang, T. M.; BMC Bioinformatics 2008, 9, 394.
9. Wang, J.; Zheng, X. Q. Math Biosci 2008, 215, 78.
37. Egan, J. P. Signal Detection Theory and ROC-Analysis; Academic Press: New York, 1975.
38. Bradley, A. P. Pattern Recogn 1997, 30, 1145.
39. Green, R. E.; Brenner, S. E. Proc IEEE 2002, 90, 1834.
40. Dai, Q.; Yang, Y. C.; Wang, T. M. Bioinformatics 2008, 24, 2296.
41. Van Helden, J. Bioinformatics 2004, 20, 399.
42. Rota, P. A.; Oberste, M. S.; Monroe, S. S.; Nix, W. A.; Campagnoli, R.; Icenogle, J. P.; Pecharanda, S.; Bankamp, B.; Maher, K.; Chen, M. H.; Tong, S.; Tamin, A.; Lowe, L.; DeRisi, J. L.; Chen, Q.; Wang, D.; Erdman, D. D.; Peret, T. C.; Burns, C.; Ksiaszek, T. G.; Rollin, P. E.; Sanchez, A.; Liffick, S.; Holloway, B.; Limor, J.; McCautland, K.; Olsen-Rasmussen, M.; Fouchier, R.; Günther, S.; Osterhaus, A. D.; Drosten, C.; Pallansch, M. A.; Anderson, L. J.; Bellini, W. J. Science 2003, 300, 1394.
43. Grigoriev, A. Trends Genet 2004, 20, 131.
44. Gu, W.; Zhou, T.; Ma, J.; Sun, X.; Lu, Z. Virus Res 2004, 101, 155.
45. Zheng, W. X.; Chen, L. L.; Ou, H. Y.; Gao, F.; Zhang, C. T. Mol Phylogenet Evol 2005, 36, 224.
46. Lai, M. M. C.; Holmes, K. V. Fields Virol 2001, 1, 1163.
47. Felsenstein, J. Cladistics 1989, 5, 164.