Integrating Overlapping Structures and Background Information of Words Significantly Improves Biological Sequence Comparison

Qi Dai[1]*, Lihua Li[2], Xiaoqing Liu[3], Yuhua Yao[1], Fukun Zhao[1], Michael Zhang[4]

[1] College of Life Sciences, Zhejiang Sci-Tech University, Hangzhou, People's Republic of China, [2] Institute for Biomedical Engineering and Instrumentation, Hangzhou Dianzi University, Hangzhou, People's Republic of China, [3] School of Science, Hangzhou Dianzi University, Hangzhou, People's Republic of China, [4] Cold Spring Harbor Laboratory, Cold Spring Harbor, New York

Abstract

Word-based models have achieved promising results in sequence comparison. However, as the important statistical properties of words in biological sequence, how to use the overlapping structures and background information of the words to improve sequence comparison is still a problem. This paper proposed a new statistical method that integrates the overlapping structures and the background information of the words in biological sequences. To assess the effectiveness of this integration for sequence comparison, two sets of evaluation experiments were taken to test the proposed model. The first one, performed via receiver operating curve analysis, is the application of proposed method in discrimination between functionally related regulatory sequences and unrelated sequences, intron and exon. The second experiment is to evaluate the performance of the proposed method with f-measure for clustering Hepatitis E virus genotypes. It was demonstrated that the proposed method integrating the overlapping structures and the background information of words significantly improves biological sequence comparison and outperforms the existing models.

Citation: Dai Q, Li L, Liu X, Yao Y, Zhao F, et al. (2011) Integrating Overlapping Structures and Background Information of Words Significantly Improves Biological Sequence Comparison. PLoS ONE 6(11): e26779. doi:10.1371/journal.pone.0026779

Editor: Dov Joseph Stekel, University of Nottingham, United Kingdom

Received May 10, 2011; **Accepted** October 4, 2011; **Published** November 10, 2011

Funding: This work is supported by the National Natural Science Foundation of China (61170316, 61001214, 61003191), National Distinguished Young Research Scientist Award (60788101), and a research grant (Y2100930) from Zhejiang Provincial Natural Science Foundation of China. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: daiailiu2004@yahoo.com.cn

Introduction

With the development of high-throughput sequencing technology, the rate of addition of new sequences to the databases increases continuously. However, such a collection of sequences does not by itself increase the scientist's understanding of the biology of organisms. Comparing a new sequence with the sequences of known functions is an effective way of assigning function to the new genes/proteins and understanding the biology of that organism from which the new sequence comes.

Owing to the importance of sequence comparison, numerous researches have been taken in past and obtained some effective tools for similarity search [1–8], evolutionary study [9–19], and classification [20–23]. The methods developed for sequence comparison can be categorized into two classes. One is alignment-based methods, in which a matrix of numbers that represents all possible alignments between two sequences is obtained with dynamic programming, and the highest set of sequential scores in the matrix defines an optimal alignment. Waterman (1995) and Durbin et al. (1998) provided comprehensive reviews about this method [24,25]. But the search for optimal solutions using alignment-based method has problems in: (i) computational load with regard to large databases [2]; (ii) choice of the scoring schemes [26]. Therefore, the emergence of research into the second class, alignment-free method, is apparent and necessary to overcome critical limitations of alignment-based methods [2,3,5,6,12,13].

Up to now, many efficient alignment-free methods have been proposed, but they are still in the early development compared with alignment-based measure [2,5,6,26–36]. One of the most widely used alignment-free approaches is word-based model that meets the need for rapid sequence comparison. In this model, each sequence is first mapped into an m-dimensional vector according to its k-word frequencies, and sequence similarity can then be measured by distance measures, such as Euclidean distance [27], Mahalanobis distance [28], Kullback-Leibler discrepancy [29,30] and Cosine distance [31]. When the k-words occurring in biological sequence are estimative probabilities rather than the frequencies, they are more readily optimized by more complex models, such as Markov model [2,33–35], mixed model [5,6] and Bernoulli model [36]. These complex models could be considered to be the modification of traditional word-based models, in which several critical problems still exist in their development as described below.

First, little attention has been paid to the overlapping structures of the words in biological sequences [2,5,27–29,31,33,34]. Overlapping occurrences of a word w are the occurrences of the word w that overlaps the previous occurrence of the word w. For instance, in the sequence ACGAATAATAAATAAGGCAATAAC, there are four occurrences of AATAA (starting at positions 4, 7, 11 and 19). But the occurrence of AATAA starting at the position 4 is different from the one starting at the position 19, because the form is composed of three overlapping occurrences of AATAA whereas the second one is composed of a unique occurrence. Because the overlapping structure of the words usually form conservative patterns in biological sequences that are strongly associated with genes [37,38], the overlapping structures of the words should be taken into account when comparing two biological sequences.

Second, background information of the words has not been fully utilized in existing biological sequence comparison [27–29,31,33,34,36]. Mutations take place randomly at molecular level, and natural selections shape the direction of evolution. In order to highlight the contribution of selective evolution, random background from the simple counting result was proposed to build a composition vector (CV) and has been used with minor modification for phylogenetic studies of prokaryotes and viruses [33,34]. Recently, Lu *et al.* found some statistical problems associated with composition vector (CV) and proposed an improved composition vector (ICV) method based on a known word distribution [36]. However, due to the fact that the word distribution is usually unknown in most cases, and each biological sequence has its own word distribution, the ICV method is of limited use.

This paper proposed an efficient statistical method for sequence comparison. It takes into consideration the overlapping occurrences of the words and has the ability to adjust the background information of the words in biological sequences. The contents can be summarized as follows:

1. An efficient word-based statistical measure based on the statistical model proposed by Schbath [39] was proposed, which utilizes the Markov model to estimate the variance of word frequencies and decomposes the similarity score into a sum of similarities of the normalized word frequencies.
2. Extensive experiments were taken to evaluate the performance of proposed model in discrimination between (a) functionally related regulatory sequences and unrelated sequences, intron and exon, and (b) different HEV genotypes. A comparison of proposed method with existing alignment-based and alignment-free models was also taken to assess its superiority.

Methods

Word-based Statistical Models (WSM)

Background information of words. A biological sequence can be described as a succession of symbols, and a k-word is a series of k consecutive letters in the sequence. For a sequence $s = s_1 s_2 \cdots s_n$, the count of a k-word $w_k = w_{k,1} w_{k,2} \cdots w_{k,k}$, denoted by $c(w_k)$, is the number of occurrence of the word w_k in the sequence s. The position of an occurrence of the word w_k is defined by the position of its first letter $w_{k,1}$. We define a random indicator $Y_i(w_k)$ of an occurrence of w_k at position i, $1 \le i \le n-k+1$, in s by

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

The occurrence frequency of the word w_k in the sequence s can be calculated with the random indicators of occurrence

$$f(w_k) = \frac{c(w_k)}{n-k+1} = \frac{\sum_{i=1}^{n-k+1} Y_i(w_k)}{n-k+1}. \qquad (1)$$

DNA and protein sequences have been realized to be a mixture of local regions that consist of compositional characteristics and pseudo-periodic sequence patterns. To utilize the background information of these local regions, we choose Markov model as a background model. It takes into consideration this 'periodical' behavior of the bio-signal by making use of transition probability matrix p and initial state distribution π.

Because $Y_i(w_k)$ is a random Bernoulli variable, the probability $\mathbb{P}(Y_i(w_k)=1)$ under the Markov model with order 1 (M^1) can be calculated by

$$\mathbb{P}(Y_i(w_k)=1|\mathrm{M}^1) = \pi(w_{k,1}) \prod_{j=2}^{k} p(w_{k,j-1}, w_{k,j}). \qquad (2)$$

For convenience, let $\mu(w_k)$ denote the probability of the word w_k to appear at a given position in the sequence, and expectation of the $Y_i(w_k)$ under the Markov model (M^1) is $\mathbb{E}[Y_i(w_k)|\mathrm{M}^1]) = \mu(w_k)$. With the expectation $\mathbb{E}[Y_i(w_k)|\mathrm{M}^1]$, we can get the expectation of the word frequency $f(w_k)$ under the Markov model (M^1)

$$\mathbb{E}[f(w_k)|\mathrm{M}^1] = \frac{\mathbb{E}[c(w_k)|\mathrm{M}^1]}{n-k+1} = \mu(w_k). \qquad (3)$$

Overlapping structures of words. Occurrences of the same word may overlap, and these overlapped words usually form a conservative pattern that is strongly associated with conservative motif [38]. So it is valuable that the overlapping structures of the words are taken into consideration when comparing two biological sequences. Here, we measure the ability of a word to overlap itself with a overlapping indicator, $\varepsilon_m(w_k)$, defined as follows:

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

where $1 \le m \le k$. With the $\varepsilon_m(w_k)$, we can calculate the probability of observing two overlapping occurrences with $k-d$ $(1 \le d \le k-1)$ letters in common and two non-overlapping occurrences of the word w_k separated by $d-k$ letters $(d \ge k)$ under the Markov model (M^1) as follows:

$$\mathbb{P}\big(Y_i(w_k)=1, Y_{i+d}(w_k)=1|\mathrm{M}^1\big) = \begin{cases} \mu(w_k)\varepsilon_{k-d}(w_k)\prod_{j=k-d+1}^{k} p\big(w_{j-1}, w_j\big) & \text{if } 1 \le d \le k, \\ \frac{[\mu(w_k)]^2}{\pi\left(w_{k,1}\right)}[p(w_{k,k}, w_{k,1})]^{d-k+1} & \text{if } d \ge k. \end{cases} \qquad (4)$$

Since the variables $Y_i(w_k)$ and $Y_{i+d}(w_k)$ are not independent under the Markov model [39–41], their effects can be described by their covariance

$$\mathbb{C}\mathrm{ov}[Y_i(w_k), Y_{i+d}(w_k)|\mathrm{M}^1] = \begin{cases} \mu(w_k)\varepsilon_{k-d}(w_k)\prod_{j=k-d+1}^{k} p(w_{k,j-1}, w_{k,j}) - \mu(w_k)^2 & \text{if } 1 \le d \le k, \\ \mu(w_k)^2\left(\frac{[p(w_{k,k}, w_{k,1})]^{d-k+1}}{\pi(w_{k,1})} - 1\right) & \text{if } d \ge k. \end{cases} \qquad (5)$$

With the above formulas, we can calculate the variance of the k-word frequency $f(w_k)$ under the Markov model (M^1)

$$\mathbb{V}_1[f(w_k)|\mathbf{M}^1] = ((n-k+1)\mu(w_k)(1-\mu(w_k)) + 2\sum_{d=1}^{k-1}(n-d-k+1)\mu(w_k)(\varepsilon_{k-d}(w_k)\prod_{j=k-d+1}^{k} p(w_{k,j-1},w_{k,j}) - \mu(w_k)) + 2\mu(w_k)^2\sum_{t=1}^{n-2k+1}(n-2k-t+2)(\frac{[p(w_{k,k},w_{k,1})]^t}{\pi(w_{k,1})} - 1))/(n-k+1)^2. \tag{6}$$

What we have presented above is the 1-order Markov model, generalizations to high order can be deduced similarly.

Word statistical model. By incorporating the overlapping structures and the background information of the words in the existing statistical model, a novel word-based statistical model is proposed and denoted in a compact form

$$\mathrm{WSM} = \{f(w_k), \mathbb{E}[f(w_k)|\mathrm{M}], \mathbb{V}\mathrm{ar}[f(w_k)|\mathrm{M}]\}. \tag{7}$$

in which the sequence information obtained through the statistical properties of the words was integrated with the overlapping structures and the background information of the words.

There are several distinctive features of this model. First, it emphasizes the structures of the words and indicates differences in terms of their contribution to the conservative patterns. Second, the influence of two overlapping occurrences of the word w_k with $k-d$ $(1 \leq d \leq k-1)$ letters in common and two non-overlapping occurrences of the word w_k separated by $d-k$ letters $(d \geq k)$ is considered. Finally, Markov model is chosen as the background model instead of Bernoulli model because each biological sequence should have its own word distribution.

Parameter estimation

Since the model parameters are priori unknown, they have to be estimated based on the observed sequences. The accuracy of this estimation is an important issue to be considered, and the existing perturbation theory for Markov chains and hidden Markov models can allow us to assess the uncertainty in the Markov chain behavior given the uncertainty [42,43]. In this paper, rather than assuming a known word distribution like [36], we estimate the model parameters with the maximum likelihood method [25] and replaces $\mathbb{E}[f(w_k)|\mathrm{M}]$ by the following estimator

$$\hat{\mathbb{E}}[f(w_k)|\mathrm{M}^r] = \frac{\prod_{j=1}^{k-r} c(w_{k,j}\cdots w_{k,j+r})}{(n-k+1)\prod_{j=2}^{k-m} c(w_{k,j}\cdots w_{k,j+r-1})}. \tag{8}$$

As for the variance, there are several approaches to derive the asymptotic variance. According to the methods proposed by Schbath [39], we have

$$\hat{\mathbb{V}}\mathrm{ar}[f(w_k)|\mathbf{M}^{k-2}] = \Big(\frac{c(w_{k,1}\cdots w_{k,k-1})c(w_{k,2}\cdots w_{k,k})}{c(w_{k,2}\cdots w_{k,k-1})^3}(c(w_{k,2}\cdots w_{k,k-1}) - c(w_{k,1}w_{k,2}\cdots w_{k,k-1})) \times (c(w_{k,2}\cdots w_{k,k-1}) - c(w_{k,2}w_{k,3}\cdots w_{k,k}))\Big)/(n-k+1)^2, k \geq 3. \tag{9}$$

However, in an application where $k \leq 2$, we derive the asymptotic variance under Markov model M^0 (Bernoulli model)

$$\hat{\mathbb{V}}\mathrm{ar}[f(w_k)|\mathbf{M}^0] = ((n-k+1)\hat{\mu}(w_k)(1-\hat{\mu}(w_k)) + 2\hat{\mu}(w_k)(\sum_{d=1}^{k-1}\varepsilon_{k-d}(w_k)\prod_{j=k-d+1}^{k}\hat{\pi}(w_{k,j}) - (k-1)\hat{\mu}(w_k)))/(n-k+1)^2, \tag{10}$$

where $\hat{\mu}(w_k)$ is the estimator of $\mu(w_k)$, $\hat{\pi}(w_{k,j})$ is the estimator of $\pi(w_{k,j})$.

Statistical similarity measure

With the assumption of the uniform distribution (U), Lu [36] calculated the word expectation and variance, and defined the normalization function ICV as:

$$\frac{f(w_k) - \hat{\mathbb{E}}[f(w_k)|\mathrm{U}]}{\sqrt{\hat{\mathbb{V}}\mathrm{ar}[f(w_k)|\mathrm{U}]}} \tag{11}$$

where $\hat{\mathbb{E}}[f(w_k)|\mathrm{U}]$ and $\hat{\mathbb{V}}\mathrm{ar}[f(w_k)|\mathrm{U}]$ are the expectation and variance of the word frequency $f(w_k)$. The normalization function ICV is necessary but not sufficient, because much effort of this method is to find better ways to utilize evolution information. In addition, the function ICV relies heavily on the word distribution. When the expectation based on background model is strongly associated with the k-word frequencies, this function can carry more information, otherwise it will increase the noise accompanied by words with exceptional background frequencies.

For the probability distributions P and Q of a discrete random variable, the relative entropy (also called Kullback-Leibler divergence) of Q from P is defined as

$$D_{KL}(P||Q) = \sum_i P(i)\log\frac{P(i)}{Q(i)} = -\sum_i P(i)\log Q(i) + \sum_i P(i)\log P(i) = H(P,Q) - H(P), \tag{12}$$

where $H(P,Q)$ is the cross entropy of P and Q, and $H(P)$ is the entropy of P. The relative entropy is the most important concept in both statistical biology and information theory. It has been deployed as non-distance similarity measures, such as kld [29,30] and $SimMM$ [2], to compare biological sequences.

A statistical measure between two proposed statistical models was proposed here based on the cross entropy $H(P,Q)$ and Euclidean distance. It is denoted by $WSMm.k.r$ as follows:

$$WSMm.k.r(WSM_X^r, WSM_Y^r) = \sum_{w^k \in S^k}(\frac{f^X(w_k)}{\hat{\mathbb{V}}\mathrm{ar}[f^X(w_k)|M_X^r]}\log(\frac{f^Y(w_k)}{\hat{\mathbb{V}}\mathrm{ar}[f^Y(w_k)|M_Y^r]}) - \frac{f^Y(w_k)}{\hat{\mathbb{V}}\mathrm{ar}[f^Y(w_k)|M_Y^r]}\log(\frac{f^X(w_k)}{\hat{\mathbb{V}}\mathrm{ar}[f^X(w_k)|M_X^r]}))^2, \tag{13}$$

where WSM_X^r and WSM_Y^r are two statistical models with Markov order r for two biological sequences X and Y, and the set S_k consists of all possible sequences of length k with symbol from the alphabet $\mathcal{A}$. In the context of DNA sequences, $\mathcal{A}$ is {A,C,G,T}. It is noticed that the similarity measure $WSMm.k.r$ satisfies the identity and triangle, but it does not satisfies inequality conditions. So it is only a dissimilarity measure. Another point of interest about this similarity measure is its normalization function that can reduce the noise by ignoring the word expectation in its definition.

Receiver operating curve and F-measure

Receiver Operating Curve analysis. Receiver operating curve (ROC) analysis has been widely used in signal detection and classification [44]. It is usually employed in binary classification of continuous data categorized as positive (1) or negative (0) cases. The classification accuracy can be measured by sensitivity and specificity, which are defined as

$$\begin{aligned} \text{sensitivity} &= \frac{\text{True Positives}}{\text{Positives}} = \frac{\text{TP}}{\text{TP}+\text{FN}}, \\ \text{specificity} &= \frac{\text{True Negatives}}{\text{Negatives}} = \frac{\text{TN}}{\text{TN}+\text{FP}}, \\ 1-\text{specificity} &= \frac{\text{FP}}{\text{TN}+\text{FP}}. \end{aligned} \qquad (14)$$

ROC curve is a graphical plot of sensitivity versus (1-specificity) for different threshold values. The area under a ROC curve (AUC) is an important value used to quantify the quality of a classification because it is a threshold independent performance measure and is closely related to the Wilcoxon signed-rank test [45]. A comprehensive discussion on AUC measure can be found in [46].

F-measure. F-measure is a measure of a test's accuracy and often used in the field of information retrieval for measuring search, document classification, and query classification performance [47]. Both the precision p and the recall r of the test are used to compute it. Here p is the number of correct results divided by the number of all returned results while r is the number of correct results divided by the number of results that should have been returned. The traditional F-measure is the harmonic mean of precision and recall:

$$\text{F} = \frac{2pr}{p+r}. \qquad (15)$$

The F-measure can be interpreted as a weighted average of the precision and recall. It ranges from 0 for highest dissimilarity to 1 for identical classifications.

Results

Evaluation on functionally related regulatory sequences

Regulatory sequence comparison plays an important role in the *abinitio* discovery of $cis-regulatory$ modules (CRMs) with a common function. If a set of co-regulated genes in a single species is given, we wish to find, in their upstream and downstream regions (henceforth called the 'control regions'), the CRMs that mediate the common aspect of their expression profiles. The control regions may be tens of Kilobase long for each gene (especially for metazoan genomes), while the CRMs to be discovered are often only hundreds of base pair long. One must therefore search in the control regions for subsequences (the candidate CRMs) that share some functional similarity [5,6].

The proposed WSM model is tested to evaluate if functionally related sequence pairs are scored better than unrelated pairs of sequences randomly chosen from the genome. In order to facilitate comparison, we choose following seven data sets published by Kantorovitz MR et al. [6]: FLY BLASTODERM (82 CRMs with expression in the blastoderm-stage embryo of the fruitfly, Drosophila 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 1220 bp) involved in regulation of the tracheal system in the fruitfly]; FLYEYE [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 by [5,6,48].

Experimental program is designed according to following settings: (1) A set of CRMs, known to regulate expression in the same tissue, is taken as the 'positive' set for each sequence in this set is the really $cis-regulatory$ module, and a set of equally many randomly chosen noncoding sequences, with lengths matching the CRMs, is taken as the 'negative' set for each sequence in this set is the randomly chosen noncoding sequence not the really $cis-regulatory$ module. It would be interesting if we choose negative sequences from nearby regions of the known CRMs (positives), which will presumably have similar word distributions. Here, we chose seven noncoding data sets published by Kantorovitz MR et al. [6] to facilitate comparison with their results. (2) Each pair of sequences in the positive set is compared, and so is each pair in the negative set. (3) 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 both from the positive and negative sets constitute a vector of length $2\binom{2}{n}$. In addition, we can get a vector of length $2\binom{2}{n}$ consisting of 1 and 0 as class labels. A perfect measure would completely separate the negative from the 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.

For comparison purpose, widely-used alignment tools were tested. These alignment tools include Needleman-Wunsch (global alignment) and 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. We also implemented four word-based measures: Euclidean distance ($eu.k$) [27], Cosine distance ($cos.k$) [31], Pearson's correlation coefficient ($pcc.k$) [32] and Kullback-Leibler discrepancy ($kld.k$) [29]. The performance of the proposed model was also compared with Markov models ($SimMM$ [2], composition vector ($CV.k.r$ [33,34]), $D.k.r$ [35]) and mixed models ($D2.k.r$ [49], $D2z.k.r$ [6], $S1.k.r$ [5] and $S2.k.r$ [5]). In addition to the alignment and statistical models, the improved composition vector ($ICV.k$) [36] was also tested. All statistical models based on the k-word distribution run with k from 2 to 8. The $CV.k.r$, $D.k.r$, $D2.k.r$, $D2z.k.r$, $S1.k.r$, $S2.k.r$ and $WSMm.k.r$ run with Markov order r from 0 to 6 and the word length k from 2 to 7. For each method, separate tests were performed with all combinations of parameter values, and the best combination was chosen to represent that score in the performance.

The AUCs for different methods are presented in Figure 1 and Table S1 in supplementary material. The first observation is that high accuracy of prediction can be achieved by the proposed measure $WSMm$. In the BLASTODERM experiment, the proposed measure $WSMm$ performs better than other alignment-based or alignment-free methods, with the area under ROC curve 0.9036. The next best method is the composition vector CV. In the PNS experiment, the measure $WSMm$ is better than all

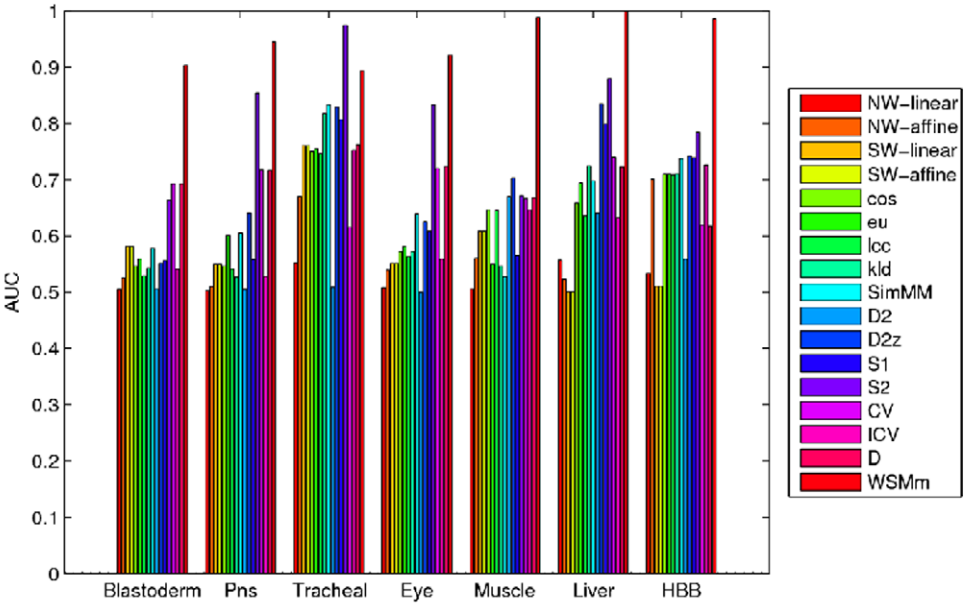

Figure 1. Comparison of AUCs of all models for detection of functionally related regulatory sequences. Comparison of AUCs of all models for detection of functionally related regulatory sequences. NW-linear and NW-affine denote Needleman-Wunsch (global alignment) raw scores, using linear gap penalties and affine gap penalties, respectively; SW-linear and SW-affine denote Smith-Waterman (local alignment) raw scores, using linear gap penalties and affine gap penalties, respectively; Word-based models are *eu*, *cos*, *pcc*, *kld*; Markov models are *SimM M*, *CV*, *D*; Mixed models are *D2*, *D2z*, *S1* and *S2*; Bernoulli model is *ICV*.
doi:10.1371/journal.pone.0026779.g001

other measures, its area under ROC curve is 0.9456. In the TRACHEAL experiment, $S2$ outperforms other measures, and its AUC is 0.975. It is followed by the measure $WSMm$. In the EYE experiment, the area under ROC curve of the measure $WSMm$ is 0.9216 , significantly better than that of other statistical methods. The next best measures is the measure $S2$. In the MUSCLE experiment, the measure $WSMm$ significantly outperforms other methods, and its area under ROC curve is 0.9892. It is followed by the $D2z$. In LIVER experiments, the measure $WSMm$ performs significantly better than other measures, with the area under ROC curve 0.9992. The next best measure is the measure $S2$. In HBB experiments, the measure $WSMm$ achieves the best performance, followed by the $S2$. From the seven experiments, we can see that the proposed measure $WSMm$ performs significantly better than other measures among six experiments, with AUC from 0.8935 to 0.9992.

Human exons and introns classification

Numerous statistical algorithms have been proposed for exons and introns classification [50–53]. A basic assumption of these algorithms is that every exon in a genome should has some distinct sequence features or properties that can distinguish it from the surrounding regions, such as introns or intergenic regions. Competitive results have been obtained in the recognition of the exons and introns of prokaryotes gene, but the discrimination of the exons and introns in human is still a difficult problem because of their limited average length.

The secondary test of the proposed model is to discriminate the human exons and introns. These data sets were organized as follows: 1200 human exons and 1200 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.

We took the previous evaluation procedure in this experiment, which make it easier to see effectiveness of various methods. The only difference lies in the parameter selection. Here all the models based on the k-word frequency run with the word length k from 2 to 6, and the $CV.k.r$, $D.k.r$, $D2.k.r$, $D2z.k.r$, $S1.k.r$, $S2.k.r$ and $WSMm.k.r$ run with Markov order r from 0 to 5 and the word length k from 2 to 6. The AUCs for different methods are presented in Figure 2 and Table S2 in supplementary material.

In terms of the discriminative power, the proposed $WSMm$ achieves the best performance compared to the existing methods, with AUC value ranging from 0.9704 to 0.9887 for the four classification tasks. These are excellent values, given that a perfect classification has an AUC score of 1, which indicates that the WSM method is very effective to distinguish exons and introns in humans in despite of their limited average length.

Clustering HEV genotype

Hepatitis E virus (HEV) is a major cause of enterically transmitted acute hepatitis in developing countries. HEV was classified recently as the sole member of the genus Hepevirus in the family Hepeviridae. Its genome consists of a single-stranded, positive-sense RNA of approximately 7.2 kb, with three partially overlapping open reading frames (ORFs: ORF1, ORF2, and ORF3). Although only one serotype has been identified to-date, HEV displays considerable genetic diversity. Based on the extensive full-length genomic variability noted among different strains, HEV has been classified into four major genotypes [54]. Here, a total of 48 full-length HEV genome sequences are retrieved from NCBI (http://www.ncbi.nlm.nih.gov/), which have been clustered into four genotypes [55–58]. Detail information on 48 full-length HEV genome sequences can be found in Table S3 in supplementary material.

This experiment aims at assessing how well the proposed model performs on identifying HEV genotype. In relation to the clustering literature [59], neighbor-joining [60] can be considered as a hierarchical method. It is chosen to clustering HEV genotypes, which is implemented in BioPerl [61]. As HEV

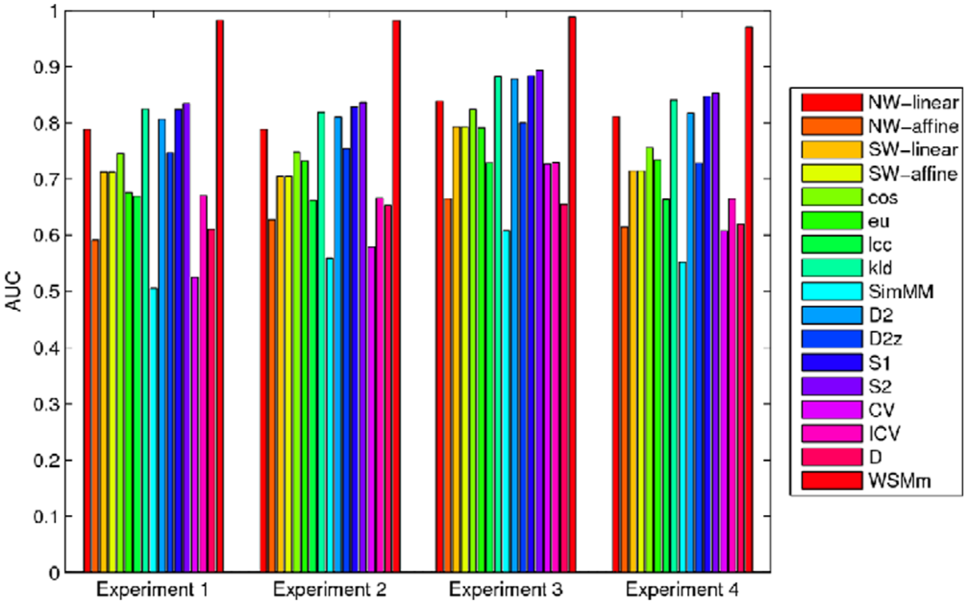

Figure 2. Comparison of AUCs of all models for classification of human exons and introns. Comparison of AUCs of all models for classification of human exons and introns. NW-linear and NW-affine denote Needleman-Wunsch (global alignment) raw scores, using linear gap penalties and affine gap penalties, respectively; SW-linear and SW-affine denote Smith-Waterman (local alignment) raw scores, using linear gap penalties and affine gap penalties, respectively; Word-based models are *eu*, *cos*, *pcc*, *kld*; Markov models are *SimM M*, *CV*, *D*; Mixed models are *D2*, *D2z*, *S1* and *S2*; Bernoulli model is *ICV*.
doi:10.1371/journal.pone.0026779.g002

genotypes is a 4-classification problem rather than one, F-measure was used to capture overall performance on HEV genotypes. To evaluate a clustering problem using the F-measure, we need to select a gold standard [59]. Here, the traditional classification was used as the gold standard [54].

In addition to the proposed method, four other typical methods were used for comparison. The used alignment-based method is Clustal W rather than Needleman-Wunsch (global alignment) or Smith-Waterman (local alignment) raw scores, because the length of genome of the HEV is approximately 7.2 kb that is difficult to handle by dynamic algorithm. The measures $D2.k.r$ and $D2z.k.r$ were not evaluated as they do not satisfy the identity condition. All statistical models based on the k-word distribution run with k from 2 to 8. The $CV.k.r$, $D.k.r$, $S1.k.r$, $S2.k.r$ and $WSMm.k.r$ run Markov order r from 0 to 7 and the word length k from 2 to 8. Figure 3 reports the F-measure for all methods on the 48 HEV genomes data set, and more details can be found in Table S4 in supplementary material.

Figure 3 shows that the proposed $WSMm.k.r$ performs better than the other alignment-based or alignment-free methods, with the F-measure 0.9791. This result is consistent with the above results, and we attribute this to the combination of both the words' overlapping structures and words' background information.

Influence of the overlapping structures of the words

For a better understanding of the proposed method, an evaluation of the word overlapping structures in biological sequences was performed. A measure, $WSMmf$, which is similar

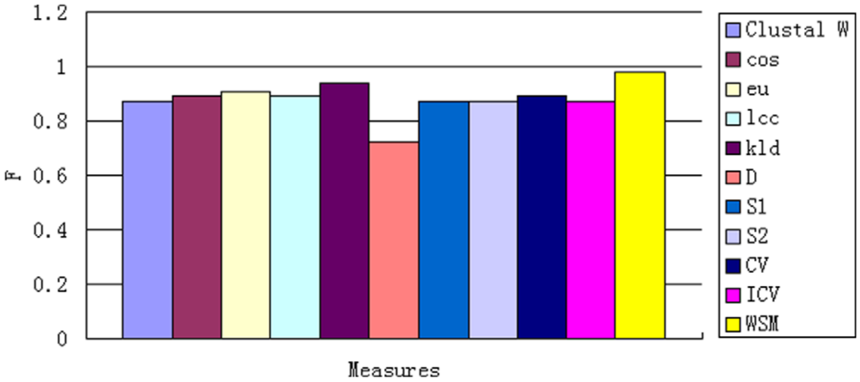

Figure 3. Comparison of F-measures of all models for classification of HEV genotypes. Comparison of F-measures of all models for classification of HEV genotypes. NW-linear and NW-affine denote Needleman-Wunsch (global alignment) raw scores, using linear gap penalties and affine gap penalties, respectively; SW-linear and SW-affine denote Smith-Waterman (local alignment) raw scores, using linear gap penalties and affine gap penalties, respectively.
doi:10.1371/journal.pone.0026779.g003

to $WSMm$ but defined based on the k-word frequencies is defined as follows:

$$WSMmf.k(\mathrm{X,Y}) = \sqrt{\sum_{w^k \in S^k} (f^X(w_k)\log f^Y(w_k) - f^Y(w_k)\log f^X(w_k))^2}, \quad (16)$$

where $f^X(w_k)$ and $f^Y(w_k)$ are the frequencies of the k-words in the biological sequences X and Y. The only difference between the measures $WSMm$ and $WSMmf$ is that the overlapping word is considered in the former. Therefore the improvement of the measure $WSMm$ can be solely attributed to the overlapping words involved. The AUCs for the measures $WSMm$ and $WSMmf$ are presented in Figure 4.

We observe that the measure $WSMm$ significantly outperforms the measure $WSMmf$ among all the experiments. For functionally related regulatory sequences, classification accuracies of the proposed measure $WSMm$ are as high as 0.8935 ~ 0.9992 in comparison to 0.5308 ~ 0.8426 with the measure $WSMmf$. For human exons and introns classification, the accuracies achieved by the proposed measure $WSMm$ is 0.9704 ~ 0.9887, while the measure $WSMmf$ only reaches 0.7871 ~ 0.8518. These results strongly demonstrate that incorporation of the overlapping words information consistently improves both efficiency and effectiveness of the sequence comparison.

Influence of the estimated word variance

Another feature of the proposed measure $WSMm$ is that the word variance is estimated upon observed biological sequences without assuming the bases occur randomly with equal chance. To show the efficiency of the estimated word variances, we compared the proposed measure $WSMm$ with another statistical measure, $WSMme$, defined as follows:

$$WSMme(\mathrm{X,Y}) = \sqrt{\sum_{i=1}^{n} (\frac{f^X(w_k)}{\mathbb{V}\mathrm{ar}[f^X(w_k)|\mathrm{E}]}\log\frac{f^Y(w_k)}{\mathbb{V}\mathrm{ar}[f^Y(w_k)|\mathrm{E}]} - \frac{f^Y(w_k)}{\mathbb{V}\mathrm{ar}[f^Y(w_k)|\mathrm{E}]}\log\frac{f^X(w_k)}{\mathbb{V}\mathrm{ar}[f_i^X|\mathrm{E}]})^2}, \quad (17)$$

where

$$\mathbb{V}\mathrm{ar}[f(w_k)|\mathrm{E}] = \frac{(\frac{n-k+1}{4^k}(1-\frac{1}{4^k}) - \frac{2}{4^{2k}}(k-1)(n-\frac{3}{2}k+1) + \frac{2}{4^k}\sum_{t=1}^{k-1}(n-k+1-t)\frac{J_t}{4^t})}{(n-k+1)^2},$$

and E denotes a known word distribution in which the four bases A, C, T, and G occur randomly with equal chance [36], k is the length of the words in biological sequences, and J_t is an indicator function, equal to 1 if $w_{k,1}\cdots w_{k,k-t} = w_{k,t+1}\cdots w_{k,k}$ and equal to 0 otherwise, for $t = 1,2,\cdots,k-1$.

The $WSMme$ assumes that the four bases A, C, T, and G occur randomly with equal chance, while the proposed measure $WSMm$ estimates the word variances according to the observed biological sequences. The comparison between the measures $WSMm$ and $WSMme$ should suggest the influence of the estimated word variance. The AUCs for the measures $WSMm$ and $WSMme$ are listed in Figure 5.

In all cases, the classification of the proposed measure $WSMm$ is more accurate than that of the measure $WSMme$. For example, by using the estimated word variance, the proposed measure $WSMm$ detects the functionally related regulatory sequences with accuracies of 0.8935 ~ 0.9992, while the measure $WSMme$ only detects 0.542 ~ 0.8426; in the case of discrimination of human exons and introns, 0.9704 ~ 0.9887 for the measure $WSMm$ contrasts with 0.8241 ~ 0.8656 for the measure $WSMme$. These results demonstrate that estimating variances from the observed sequences could be more promising to improve the biological

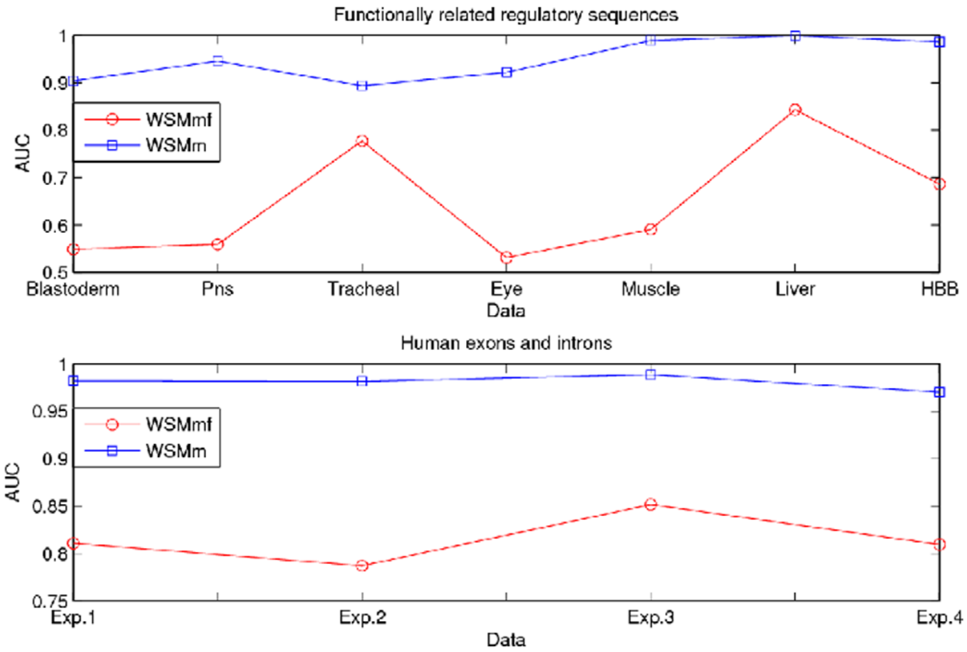

Figure 4. Comparison of AUCs of the measures *WSMm* and *WSMmf*. From top down, comparison of AUCs of the measures *WSMm* and *WSMmf* for predicting functionally related regulatory sequences and classifying human exons and introns.
doi:10.1371/journal.pone.0026779.g004

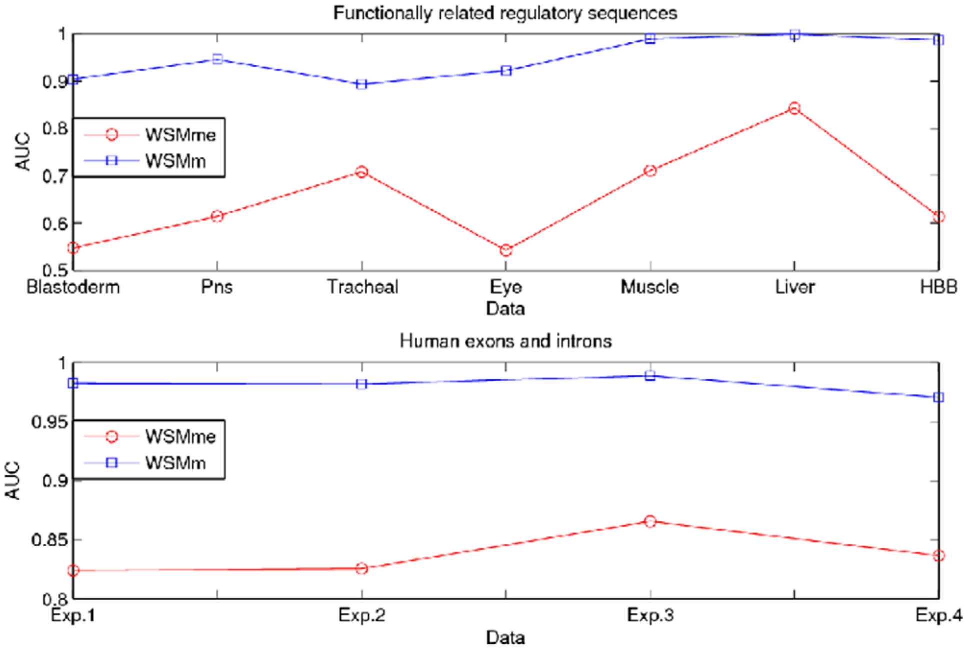

Figure 5. Comparison of AUCs of the measures *WSMm* and *WSMme*. From top down, comparison of AUCs of the measures *WSMm* and *WSMme* for predicting functionally related regulatory sequences and classifying human exons and introns.
doi:10.1371/journal.pone.0026779.g005

sequence comparison because it helps the measure $WSMm$ to adjust the background information according to the word distribution.

Discussion

This paper proposed an efficient statistical method for biological sequence comparison, which integrates both the overlapping structures and background information of the words in biological sequences. It compares biological sequence by taking advantage of the tendency of the k-word conservation. In the application, the proposed method treats the word appearing at a given position as a random variable, estimates the word variance according to the observed sequence, and therefore maximizes the impact of the overlapping structures and background information of the words in sequence. A similar idea was proposed in our previous measures $S1$ and $S2$, but as shown in our experiments, the proposed measure $WSMm$ performs significantly better which suggests that the overlapping structures and background information of the words should be included in word-based statistical methods to improve biological sequence comparison.

The proposed method originates from the existing methods but different from them in several key aspects. Blaisdell, Wu et al. and Stuart et al. [27,29,31] developed popular sequence comparison methods where similarity/dissimilarity score depends on the measure under the frequency vector of the k-words in biological sequence. However, they did not use the background information of k-words for sequence comparison, and the probability of the k-words under these models is estimated by the occurrences of the k-words. Pham and Zuegg [2] also proposed ways to improve biological sequence comparison, but their model is different from ours in that the appearance of the k-words are modeled by a Markov model, whose parameters are independent of the k-word distribution in biological sequence. We developed a Markov plus k-word distribution model [5], based on the idea of adding k-word distribution in sequence to Markov model directly. The way of treating sequence comparison is also different from the proposed method: no information about the overlapping structure of a word in biological sequence was considered in our previous mixed model. Lu et al. [36] found some statistical problems associated with composition vector (CV) [33,34] and proposed an improved composition vector (ICV) method. Their study assumes that the four bases A, C, T, and G occur randomly with equal chance and derives the expected count of a k-word and the count variance in a given sequence s based upon this simple assumption. In other words, the word distribution is assumed to be known a priori. But, in most cases the word distribution is usually unknown, and therefore the application of ICV method is very limited in practice. Most importantly, this research demonstrated that integration the overlapping structure of a word with the estimated background information of the words according to the observed sequences is essential to improve biological sequence comparison. In addition, among tree kinds of the experiments, the length of biological sequence varies from 201 (HUMAN LIVER [9 CRMs (average length 201) driving expression specific to the human liver]) to 7.2 kb (the genome of HEV consists of a single-stranded, positive-sense RNA of approximately 7.2 kb). The proposed method achieved the best performance among all the experiments, which indicates that its performance is not influenced by the sequence length. As for the computational efficiency, because the k-words in biological sequence are considered in the definition of the statistical measure $WSM.k.r$, its computational efficiency is the same as that of existing methods based on the word-based models [2,5,27–29,31,33,34,36].

One major limitation of the proposed method is that different k-words are assumed to be independent under Bernoulli and Markov model which is not always met in practice, and their influence should be taken into consideration. One consequence of our simplification is that the correlations between different k-words are ignored and only the same k-word variances are accounted for. A better model should reflect the data covariance structure. Despite of this simplification, we found that the proposed statistical measure essentially improves biological sequence comparison.

Supporting Information

Table S1 AUCs obtained from all the models for detection of functionally related regulatory sequences.
(DOC)

Table S2 AUCs obtained from all the models for classification of human exons and introns.
(DOC)

Table S3 Abbreviation for the strains, accession number, nucleotide length, genotype, and country for each of the 48 complete HEV genomes.
(DOC)

Table S4 F-measures obtained from all the models for classification of HEV genotypes.
(PDF)

Acknowledgments

The authors thank all the anonymous referees for their valuable suggestions and support. In particular, the authors thank Prof. Tuan D. Pham for providing all MATLAB code for SimMM and Prof. Saurabh Sinha for provding the software of D2z, datasets and the technical help.

Author Contributions

Conceived and designed the experiments: QD XL. Performed the experiments: QD XL. Analyzed the data: QD XL YY. Contributed reagents/materials/analysis tools: XL FZ. Wrote the paper: QD LL MZ.

References

1. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25: 3389–3402.
2. Pham TD, Zuegg J (2004) A probabilistic measure for alignment-free sequence comparison. Bioinformatics 20: 3455–3461.
3. Pham TD (2007) Spectral distortion measures for biological sequence comparisons and database searching. Pattern Recog 40: 516–529.
4. Smith AA, Vollrath A, Bradfield CA, Craven M (2008) Similarity Queries for Temporal Toxicogenomic Expression Profiles. PLoS Comput Biol 4(7): e1000116.
5. 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.
6. Kantorovitz MR, Robinson GE, Sinha S (2007) A statistical method for alignment-free comparison of regulatory sequences. Bioinformatics 23: i249–i255.
7. Van Helden J (2004) Metrics for comparing regulatory sequences on the basis of pattern counts. Bioinformatics 20: 399–406.
8. Sinha S, He X (2007) MORPH: Probabilistic alignment combined with hidden Markov models of cis-regulatory modules. PLoS Comput Biol 3(11): e216.
9. Felsenstein J (1996) Inferring phylogenies from protein sequences by parsimony, distance and like-lihood methods. Meth Enzymol 266: 418–427.
10. Huelsenbeck JP, Ronquist F (2001) MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17: 754–755.
11. Kumar S, Tamura K, Nei M (2004) MEGA3: integrated software for molecular evolutionary genetics analysis and sequence alignment. Briefing Bioinform 5: 150–163.
12. Li M, Badger JH, Chen X, Kwong S, Kearney P, et al. (2001) An information-based sequence distance and its application to whole mitochondrial genome phylogeny. Bioinformatics 17: 149–154.
13. Otu HH, Sayood K (2003) A new sequence distance measure for phylogenetic tree construction. Bioinformatics 19: 2122–2130.
14. Ronquist F, Huelsenbeck JP (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19: 1572–1574.
15. Cannarozzi G, Schneider A, Gonnet G (2007) A phylogenomic study of human, dog, and mouse. PLoS Comput Biol 3(1): e2.
16. Abeln S, Teubner C, Deane CM (2007) Using phylogeny to improve genome-wide distant homology recognition. PLoS Comput Biol 3(1): e3.
17. Rivas E, Eddy SR (2008) Probabilistic Phylogenetic Inference with Insertions and Deletions. PLoS Comput Biol 4(9): e1000172.
18. Engelen S, Trojan LA, Sacquin-Mora S, Lavery R, Carbone A (2009) Joint Evolutionary Trees: A Large-Scale Method To Predict Protein Interfaces Based on Sequence Sampling. PLoS Comput Biol 5(1): e1000267.
19. Komatsu K, Zhu S, Fushimi H, Qui TK, Cai S, et al. (2001) Phylogenetic analysis based on 18S rRNA gene and matK gene sequences of Panax vietnamensis and five related species. Planta Med 67: 461–465.
20. 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.
21. Pipenbacher P, Schliep A, Schneckener S, Schonhuth A, Schomburg D, et al. (2002) ProClust: improved clustering of protein sequences with an extended graph-based approach. Bioinformatics 18: S182–S191.
22. Levy ED, Pereira-Leal JB, Chothia C, Teichmann SA (2006) 3D Complex: A structural classification of protein complexes. PLoS Comput Biol 2(11): e155.
23. Chao KM, Zhang LX (2008) Sequence Comparison: Theory and Methods, Springer.
24. Waterman MS (1995) Introduction to Computational Biology: Maps, Sequences, and Genomes: Interdisciplinary Statistics, Chapman and Hall/CRC, Boca Raton, FL.
25. Durbin R, Eddy SR, Krogh A, Mitchison G (1998) Biological Sequence Analysis, Cambridge University Press.
26. Vinga S, Almeida J (2003) Alignment-free sequence comparison-a review. Bioinformatics 19: 513–523.
27. Blaisdell BE (1986) Ameasure of the similarity of sets of sequences not requiring sequence alignment. Proc Natl Acad Sci USA 83: 5155–5159.
28. 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.
29. Wu TJ, Hsieh YC, Li LA (2001) Statistical measures of DNA dissimilarity under Markov chain models of base composition. Biometrics 57: 441–448.
30. Ulitsky I, Burstein D, Tuller T, Chor B (2006) The average common substring approach to phylogenomic reconstruction. J Comput Biol 13: 336–350.
31. Stuart GW, Moffett K, Baker S (2002) Integrated gene and species phylogenies from unaligned whole genome protein sequences. Bioinformatics 18: 100–108.
32. Fichant G, Gautier C (1987) Statistical method for predicting protein coding regions in nucleic acid sequences. Comput Appl Biosci 3: 287–295.
33. Hao B, Qi J (2004) Prokaryote phylogeny without sequence alignment: from avoidance signature to composition distance. J Bioinform Comput Biol 2: 1–19.
34. 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.
35. Apostolico A, Denas O (2008) Fast algorithms for computing sequence distances by exhaustive substring composition. Algorithms Mol Biol 3: 13.
36. Lu GQ, Zhang SP, Fang X (2008) An improved string composition method for sequence comparison. BMC Bioinformatics 9(Suppl 6): S15.
37. Livak F (2003) Evolutionarily conserved pattern of gene segment usage within the mammalian TCRbeta locus. Immunogenetics 55: 307–314.
38. Dixon RJ, Eperon IC, Samani NJ (2007) Complementary intron sequence motifs associated with human exon repetition: a role for intragenic, inter-transcript interactions in gene expression. Bioinformatics 23: 150–155.
39. Schbath S (1997) An E±cient Statistic to Detect Over-and Under-Represented Words in DNA Sequences. J Comp Biol 4(2): 189–192.
40. Reinert G, Schbath S, Waterman MS (2000) Probabilistic and statistical properties of words: an overview. J Comput Biol 7: 1–46.
41. Robin S, Daudin JJ (1999) Exact distribution of word occurrences in a random sequence of letters. J Appl Prob 36: 179–193.
42. Mitrophanov AY (2005) Sensitivity and convergence of uniformly ergodic Markov chains. J Appl Prob 42: 1003–1014.
43. Mitrophanov AY, Lomsadze A, Borodovsky M (2005) Sensitivity of hidden Markov models. J Appl Prob 42: 632–642.
44. Egan JP (1975) Signal Detection Theory and ROC-Analysis, Academic Press, New York.
45. Bradley AP (1997) The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recog 30: 1145–1159.
46. Green RE, Brenner SE (2002) Bootstrapping and normalization for enhanced evaluations of pairwise sequence comparison. Proc IEEE 90: 1834–1847.
47. Rijsbergen CJ (1979) Information Retireval, Butterworths, London.
48. Gallo SM, Li L, Hu Z, Halfon MS (2006) REDfly: a Regulatory Element Database for Drosophila. Bioinformatics 22: 381–383.
49. Lippert RA, Huang HY, Waterman MS (2002) Distributional regimes for the number of k-word matches between two random sequences. Proc Natl Acad Sci USA 99: 13980C13989.
50. Guigo R (1999) In Genetic Databases, Academic Press, New York.
51. Wu YH, Liew AWC, Yan H, Yang MS (2003) Classification of short human exons and introns based on statistical features. PHYSICAL REVIEW E 67(6): 061916.
52. Jiang R, Yan H (2008) Segmentation of short human exons based on spectral features of double curves. IJDMB 2(1): 15–35.

53. Jiang R, Yan H (2008) Studies of spectral properties of short genes using the wavelet subspace Hilbert Huang transform(WSHHT)[J]. Physica A 387: 4223–4247.
54. 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.
55. Xia HY, Liu LH, Linde AM, Belak S, Norder H, et al. (2008) Molecular characterization and phylogenetic analysis of the complete genome of a hepatitis E virus from European swine. Virus Genes 37: 39C48.
56. Liu L, Xia H, Wahlberg N, Belok S, Baule C (2009) Phylogeny, classification and evolutionary insights into pestiviruses. Virology 385: 351C357.
57. Olvera A, Busquets N, Cortey M, de Deus N, Ganges L, et al. (2010) Applying phylogenetic analysis to viral livestock diseases: moving beyond molecular typing. Vet J 184(2): 130–137.
58. 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.
59. Handl J, Knowles J, Kell DB (2005) Computational Cluster Validation in Post-Genomic Data Analysis. Bioinformatics 21: 3201–3212.
60. Felsenstein J (1989) PHYLIP-Phylogeny inference package (version 3.2). Cladistics 5: 164–166.
61. Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, et al. (2002) The BioPerl Toolkit: Perl Modules for the Life Sciences. Genome Res 12: 1611–1618.