BMC Bioinformatics

Methodology article

Comparison study on k-word statistical measures for protein: From sequence to 'sequence space'

Qi Dai* and Tianming Wang

Address: Department of Applied Mathematics, Dalian University of Technology, Dalian 116024, PR China

Email: Qi Dai* - daiailiu2004@yahoo.com.cn; Tianming Wang - wangtm@dlut.edu.cn

* Corresponding author

Published: 23 September 2008

BMC Bioinformatics 2008, **9**:394 doi:10.1186/1471-2105-9-394

Received: 9 April 2008

Accepted: 23 September 2008

This article is available from: http://www.biomedcentral.com/1471-2105/9/394

Abstract

Background: Many proposed statistical measures can efficiently compare protein sequence to further infer protein structure, function and evolutionary information. They share the same idea of using *k*-word frequencies of protein sequences. Given a protein sequence, the information on its related protein sequences hasn't been used for protein sequence comparison until now. This paper proposed a scheme to construct protein 'sequence space' which was associated with protein sequences related to the given protein, and the performances of statistical measures were compared when they explored the information on protein 'sequence space' or not. This paper also presented two statistical measures for protein: *gre.k* (generalized relative entropy) and *gsm.k* (gapped similarity measure).

Results: We tested statistical measures based on protein 'sequence space' or not with three data sets. This not only offers the systematic and quantitative experimental assessment of these statistical measures, but also naturally complements the available comparison of statistical measures based on protein sequence. Moreover, we compared our statistical measures with alignment-based measures and the existing statistical measures. The experiments were grouped into two sets. The first one, performed via ROC (Receiver Operating Curve) analysis, aims at assessing the intrinsic ability of the statistical measures to discriminate and classify protein sequences. The second set of the experiments aims at assessing how well our measure does in phylogenetic analysis. Based on the experiments, several conclusions can be drawn and, from them, novel valuable guidelines for the use of protein 'sequence space' and statistical measures were obtained.

Conclusion: Alignment-based measures have a clear advantage when the data is high redundant. The more efficient statistical measure is the novel *gsm.k* introduced by this article, the *cos.k* followed. When the data becomes less redundant, *gre.k* proposed by us achieves a better performance, but all the other measures perform poorly on classification tasks. Almost all the statistical measures achieve improvement by exploring the information on 'sequence space' as word's length increases, especially for less redundant data. The reasonable results of phylogenetic analysis confirm that *Gdis.k* based on 'sequence space' is a reliable measure for phylogenetic analysis. In summary, our quantitative analysis verifies that exploring the information on 'sequence space' is a promising way to improve the abilities of statistical measures for protein comparison.

Background

Over the past few decades, major advances in the field of molecular biology, coupled with advances in genomic technologies, have led to an explosive growth of biological sequences databases. For example, there are several well-known databases about protein: Pfam [1] (a secondary database for multiple alignments and profile hidden Markov models), SCOP [2] (a secondary database containing protein family and structural information), Swiss-Prot [3] (primary database of protein sequences), and Protein Information Resource (PIR) [4] (primary database of protein sequences). This deluge of databases, in turn, produces new questions to analyze protein sequences such as how to classify protein sequences, induce their evolutionary information, and predict their structures.

Among protein sequence analysis, some important computational methods are similarity search, phylogenetic analysis and sequence classification. The similarity search [5-7] is to search a database of known function sequences and uses the structures and functions of the most closely matched known sequences to analyze the structure and function of query sequence. Phylogenetic analysis [8-12] is the study of the evolutionary history among species. It can also provide useful information for pharmaceutical researchers to determine which species share the medicinal qualities [13]. Classification protein [14,15] is to get a biologically meaningful partition. It has several advantages: when proteins are grouped into a family, it can provide us some clues about the general features of this family and evolutionary evidence of proteins, and further infer the biological function of a new sequence by its similarity to some function-known sequences. Moreover, protein classification can be used to facilitate protein three-dimensional structure discovery, which is very important for understanding proteins' functions. However, these computational methods heavily rely on the (dis)similarity measures defined among biological sequences.

Because of the importance of research into (dis)similarity measures, numerous efficient algorithms have been developed, but challenges remain. Moreover, we believe that further improvements in the (dis)similarity measures will allow us to design more effective tools, which can help us to look back more deeply in evolutionary time. One kind of the most common dissimilarity measures in this area is edit distance by aligning two sequences. It is defined as the required number of insertions, deletions, and replacements of characters from the first protein sequence to obtain the second protein sequence. But this measure is encountered with difficulties: (i) computation with regard to large biological databases [16,17]; (ii) the score schemes chosen [16]. Therefore, alignment-free measures are actively pursued to overcome the limitations of protein analysis by alignment.

Up to now, many efficient alignment-free measures for sequences comparison have been proposed, but they are still in the early development compared with alignment-based methods. One of the comprehensive reviews [16] reported several concepts of (dis)similarity measures, such as Euclidean distance [18], Mahalanobis distances [19], Kullback-Leibler discrepancy [20], Cosine distance [21] and Pearson's correlation coefficient [22]. Recently, several novel alignment-free measures have been designed for protein sequences analysis, such as S1 and S2 [23], W-metric [14], Universal Similarity Metric [15], Local decoding [24], CLUSS [25] and Long Short-Term Memory [26].

Among the statistical measures, each sequence is mapped into an n-dimensional vector according to its k-word frequencies. Linear Algebra theory is further employed to define the similarity score between sequences represented in vector spaces. The *kld* extended by Wu et al. (2001) is computed in terms of two vectors of relative frequencies of k-words over a sliding window from two given DNA sequences. However, in an application where some entries of vectors are equal to 0 or 1, *kld* becomes unsuitable. In this paper, we present two statistical measures to overcome the limitation of the measure *kld*. The contents can be summarized as follows:

1. We present a scheme to build protein 'sequence space' based on the score or amino acid substitution matrices and calculate k-word frequencies of protein 'sequence space'.

2. Two statistical measures *gre.k* and *gsm.k*, as the extended Jensen-Shannon Divergence, are proposed. They are based on k-word frequencies and Jensen-Shannon Divergence. Although these two concepts are not new, their generalizations result in the novel aspect of these measures. Particularly, the statistical measure, *Gdis.k*, is proved to be a valid distance measure.

3. Our measures are applied to extensive tests, e.g., protein sequence classification and phylogenetic analysis. The performances of our measures are compared with alignment-based measures and the existing statistical measures. Through the experiments, we want to address the following questions with the aid of well known statistical index: (A) how well our statistical measures perform compared with the existing statistical measures and alignment-based ones; (B) which statistical measure performs better when exploring the information on protein 'sequence space'; (C) whether the classification abilities of statistical measures depend on the choice of score matrices; (D) whether our measure, *Gdis.k*, is a valid distance measure for phylogenetic analysis.

Results and discussion

Classification of protein sequences

The proposed statistical measures are used to classify protein sequences. Several benchmark data sets of non-homologous protein structures have been developed in the last few years [27-30]. In this study, we have chosen the 36 protein domains of [27], the Rost and Sander data set (RS) and the 86 prototype protein domains of [28]. The Chew-Kedem data set (Additional file 1) was introduced in [27] and further studied in [31]. It consists of 36 protein domains drawn from PDB entries of three classes (alpha/beta, mainly-alpha, mainly-beta). Although this data set has been extensively used, the main draw back of this data is small size and high redundant. The Rost and Sander data set (RS126) (Additional file 2) was designed for the secondary structure prediction of proteins with a pair-wise sequence similarity of less than 25% [32], and it was used as a test data to evaluate the performances of similarity measures [33]. Here, we not only compare the proteins' secondary structures, but analyse the performance of (dis)similarity measures according to the proteins' classification as given by SCOP, release 1.69 [34]. We adopt this manually curated database as our gold standard containing expert knowledge for class level. This data set is trimmed to exclude sequences belonging to classes with <5 elements, thus a data set of 121 protein sequences, denoted by RS, is obtained. The Sierk-Pearson data set (Additional file 3), which consists of a non-redundant subset of 2771 protein families and 86 non-homologous protein families from the CATH protein domain database [35], was introduced in [28]. We estimate the homology of the data by employing CD-HIT program, which clusters protein databases at given sequence homology threshold [36]. Running CD-HIT with 70% homology threshold reveals that there are 29, 120, 86 sequences for data CK, RS and SP, respectively, below the homology threshold. This results clearly indicate that CK is high redundant, RS is low redundant, and SP is less redundant.

The experiments aim at evaluating the classification ability of the alignment-based measures and the statistical measures. The evaluation procedure is based on a binary classification of each protein pair, where 1 corresponds to the two protein sequences sharing the same class, 0 otherwise.

Given a data with size n, a $n \times n$ similarity/distance matrix can be obtained via each measure. The entries of the upper triangular similarity/distance matrix constitute a similarity vector of length $\binom{n}{2}$, which is used as prediction. Also, we can get a vector of length $\binom{n}{2}$ consisted 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 measures evaluated are: alignment-based measures, our statistical measures (*gre.k* and *gsm.k*) and the six statistical measures outlined in Method section (*ed.k*, *cos.k*, *se.k*, *W.k*, *s1.k* and *s2.k*), where the alignment-based measures are Clustal X, Needleman-Wunsch (global alignment) or Smith-Waterman (local alignment) raw scores, with no correction for statistical significance, using ten score matrices (BLOSUM40, BLOSUM45, BLOSUM62, BLOSUM80, BLOSUM100, PAM40, PAM80, PAM120, PAM200, PAM250) and linear gap penalties or affine gap penalties, with a gap penalty of 8. All statistical measures based on k-word frequencies of protein sequence and protein 'sequence space' run with k from 1 to 4, where protein 'sequence space' is constructed based on the score matrix (BLOSUM40, BLOSUM45, BLOSUM62, BLOSUM80, BLOSUM100, PAM40, PAM80, PAM120, PAM200, PAM250). For each measure, separate tests are done with each combination of parameter values, and the best combination is chosen to represent the score in the performance. ROC curves are computed to evaluate and compare the performances of our methods and other (dis)similarity measures.

The ROC curves obtained for the classifications are presented in Figures 1, 2, 3. Figure 1(a), Figure 2(a) and Figure 3(a) denote the ROC curves of alignment-based measures and the statistical measures based on k-word frequencies of protein sequences. Figure 1(b), Figure 2(b) and Figure 3(b) denote the ROC curves of alignment-based measures and the statistical measures based on k-word frequencies of protein 'sequence space'. The better (dis)similarity measures have plots with higher values of sensitivity for equal values of specificity, resulting in higher values for the areas under the curves. The AUC value is typically used as a measure of overall discrimination accuracy. Table 1 provides the areas under ROC curves (AUC) obtained from all the (dis)similarity measures for data sets CK, RS and SP.

Question A

In the CK experiment, Figure 1 and Table 1 show that alignment-based measures perform better than alignment-free measures. NW-affine.b45 outperforms other alignment-based measures, its area under ROC curve is

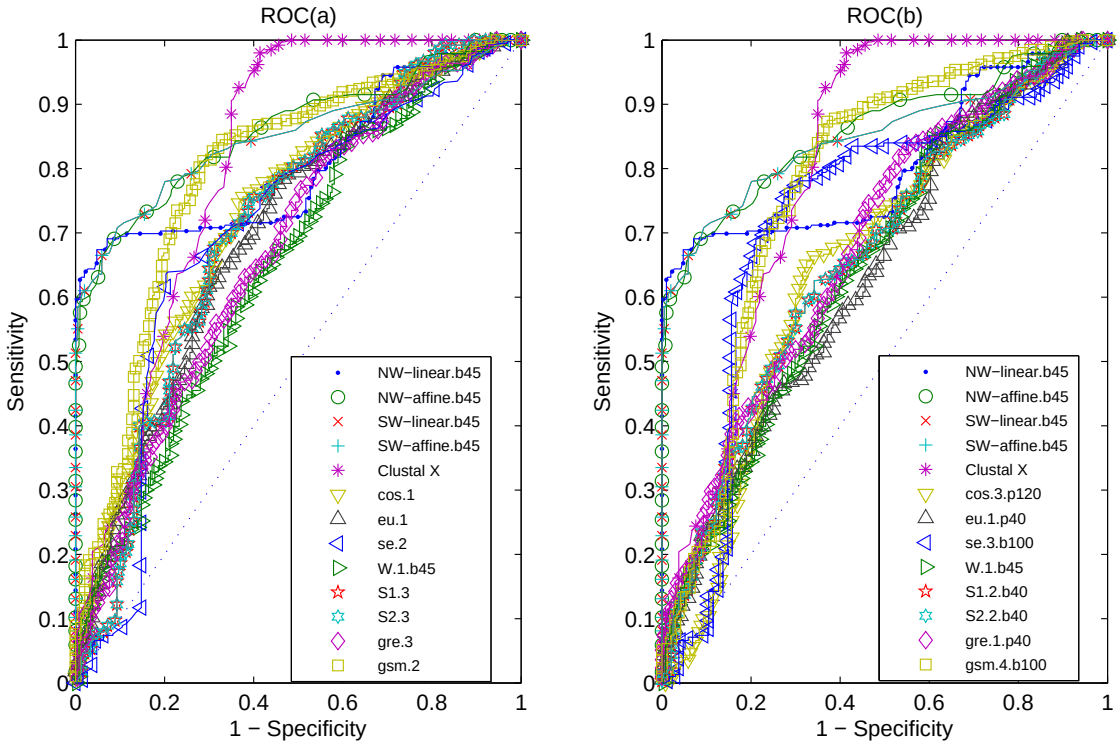

Figure 1
ROC curves for data CK. ROC (a) for our measures, alignment-based measures and other statistical measures, all the statistical measures are based on *k*-word frequencies of protein sequence, with the parameter values as suffix. ROC (b) for our measures, alignment-based measures and other statistical measures, all the statistical measures are based on *k*-word frequencies of protein 'sequence space', with the parameter values as suffix. All the abbreviations of (dis)similarity measures are illustrated in the "List of abbreviations" section. A random classifier would generate equal proportions of FP and TP classifications, which corresponds to the ROC diagonal (dashed line).

0.860. Among the statistical measures based on *k*-word frequencies of protein sequences, *gsm.2* is clearly more efficient than other measures. Its area under ROC curve is 0.791. The next best measure is the *cos.1*, with the area under ROC curve 0.729, and the other measures lag behind. For the statistical measures based on *k*-word frequencies of protein 'sequence space', *gsm.4.b100* is significantly better than other statistical measures, the *se.3.b100* followed.

In the RS experiment, Figure 2 and Table 1 indicate that some statistical measures perform as well as alignment-based measures. By exploring the information on protein 'sequence space', the statistical measure, *gsm.k*, performs better than alignment-base measures. For the alignment-based measures, *NW-affine.b40* performs better than other measures. As for the statistical measures based on *k*-word frequencies of protein sequences, *cos.1* outperforms the other measures. Among the statistical measures based on *k*-word frequencies of protein 'sequence space', *gsm.3.b40* is significantly better than all other measures, its area under ROC curve is 0.627, and the next best measure is *gre.4.b100*.

In the SP experiment, Figure 3 and Table 1 illustrate that some statistical measures defined by *k*-word frequencies of protein sequences outperform alignment-based measures. When the information on protein 'sequence space' is added, all the statistical measures, except for *se.k* and *s2.k*, perform better than alignment-base measures. For the alignment-based measures, *SW* measures perform better than *NW* measures. As for the statistical measures based on *k*-word frequencies of protein sequences, *gre.1* outperforms other measures, which is followed by cos.1 and *eu.1*. Among the statistical measures based on *k*-word frequencies of protein 'sequence space', the area under ROC curve of *gre.1.p40* is 0.575, better than other statistical measures, and the next best measures are the *cos.1.p40* and *eu.1.p40*.

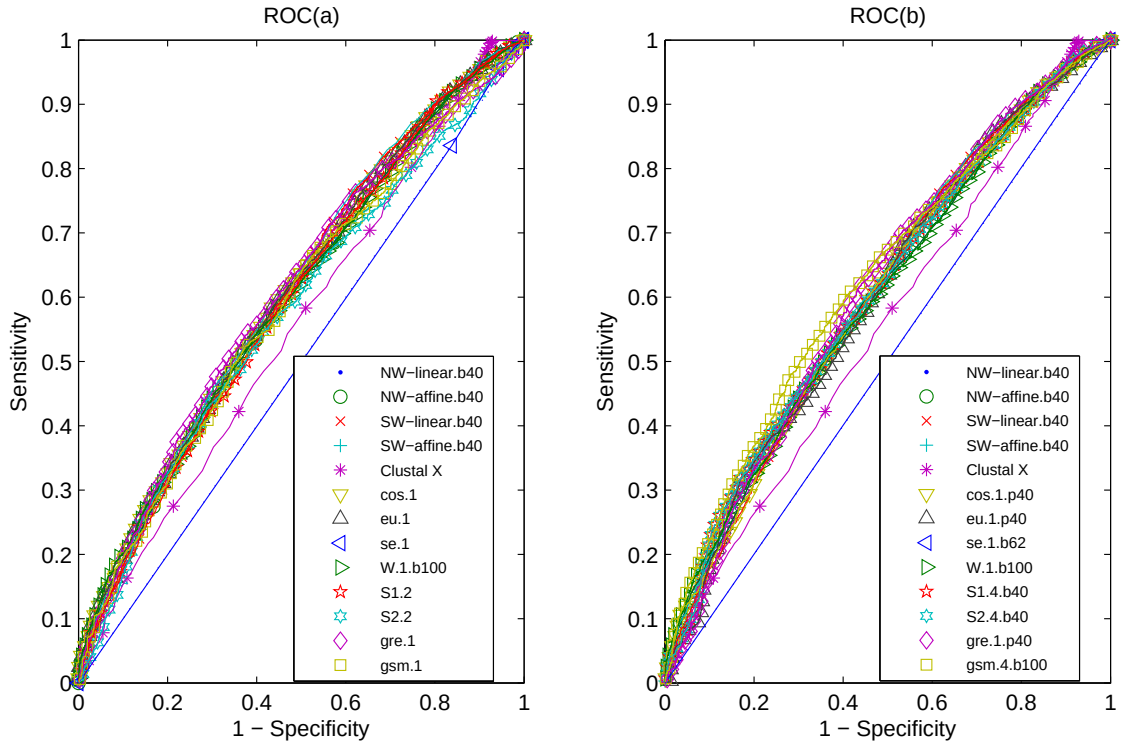

Figure 2
ROC curves for data RS. ROC (a) for our measures, alignment-based measures and other statistical measures, all the statistical measures are based on *k*-word frequencies of protein sequence, with the parameter values as suffix. ROC (b) for our measures, alignment-based measures and other statistical measures, all the statistical measures are based on *k*-word frequencies of protein 'sequence space', with the parameter values as suffix. All the abbreviations of (dis)similarity measures are illustrated in the "List of abbreviations" section. A random classifier would generate equal proportions of FP and TP classifications, which corresponds to the ROC diagonal (dashed line).

From the above three experiments, we can see that alignment-based measures have a clear advantage when the data is high redundant. The most efficient statistical measure is the novel *gsm.k* introduced by this report. When the data becomes less redundant, *gre.k* proposed by us achieves a better performance, but all the alignment-based and the existing measures perform poorly on all classification tasks. The inspection of the ROC curves themselves (Figures 1, 2, 3) further illustrates these comparisons between (dis)similarity measures.

Question B

The main goal of construction of protein 'sequence space' is to improve the classification ability of (dis)similarity measures by extracting the information on related protein sequences. However, it should be noted that not all the (dis)similarity measures are suitable for this scheme. In order to find which statistical measure is suitable for this scheme, we define a function *DAUC (measure, score matrix, k)* to evaluate whether the classification ability of (dis)similarity measures improve or not,

$$DAUC\ (measure, score\ matrix, k) = AUC\ (measure, score\ matrix, k) - AUC\ (measure, k), \quad (1)$$

where *AUC (measure, score matrix, k)* denotes the area under ROC curve of the statistical measure based on the *k*-word frequencies of protein 'sequence space', which is constructed based on the score matrix; *AUC (measure, k)* denotes the area under ROC curve of measure defined by the *k*-word frequencies of protein sequence.

Judging from definition of *DAUC*, it is easier to recognize that if DAUC ≥ 0, utilizing protein 'sequence space' improves the classification ability of the (dis)similarity measures. The *DAUC* values for the data CK, RS and SP are presented in Figures 4, 5, 6.

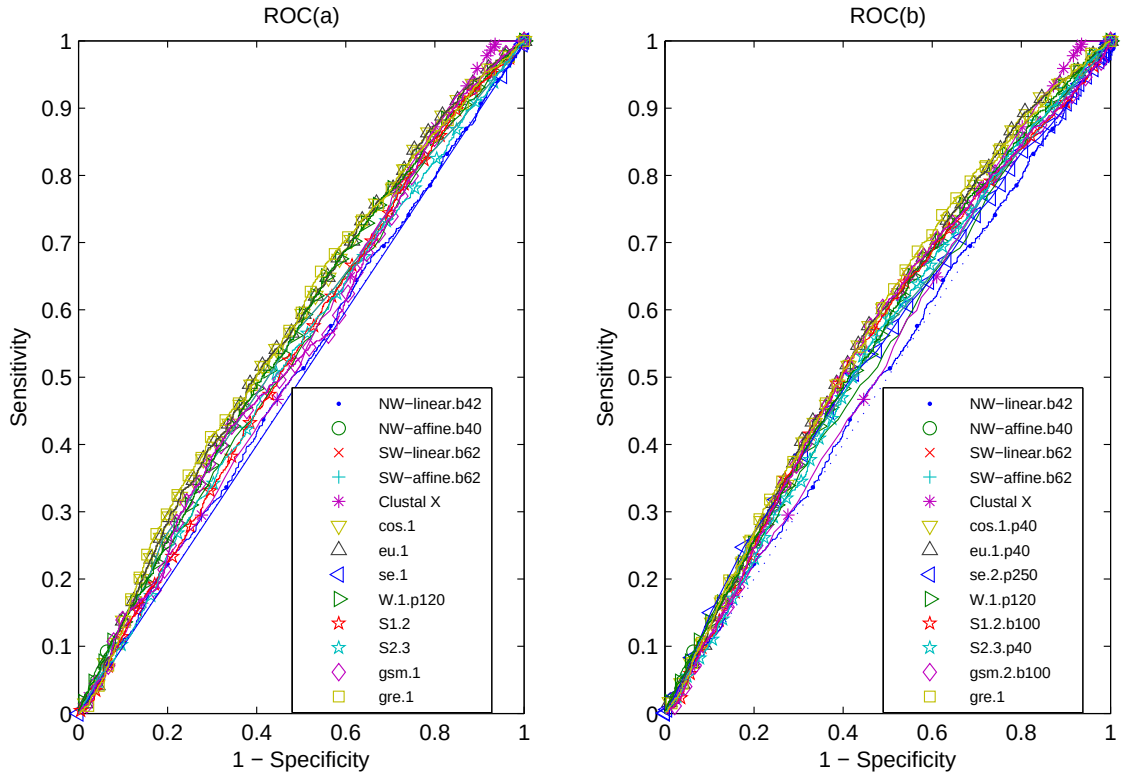

Figure 3

ROC curves for data SP. ROC (a) for our measures, alignment-based measures and other statistical measures, all the statistical measures are based on *k*-word frequencies of protein sequence, with the parameter values as suffix. ROC (b) for our measures, alignment-based measures and other statistical measures, all the statistical measures are based on *k*-word frequencies of protein 'sequence space', with the parameter values as suffix. All the abbreviations of (dis)similarity measures are illustrated in the "List of abbreviations" section. A random classifier would generate equal proportions of FP and TP classifications, which corresponds to the ROC diagonal (dashed line).

As would be expected, the *DAUC* values of the different measures (Figures 4, 5, 6) show two clear trends: (i) the *DAUC* values increase from $k = 1$ to $k = 4$ for all three data sets. When the length of word is equal to 4, almost all the statistical measures' classification abilities are improved. It should be noted that the classification discrimination of statistical measures based on higher order word frequencies, such as *eu.k*, *se.k* and *cos.k*, worsens [14], because the high dimension of the frequency vectors and the relative low dimension of the sequences length itself cause the frequency vector *F* to be very sparse. Interestingly, the construction of protein 'sequences space' maintains the accuracy and overcomes the difficulty arising from higher order word; (ii) it is interesting to note that there is a dependency between usefulness of protein 'sequence space' and the level of data's redundant. When the data is high redundant such as CK, the 'sequence space' is more similar. Consequently, the (dis)similarity measures based on 'sequence space' achieve a little improvement (Figure 4 (k = 4)). But the accuracy of classification is also improved with word's length increasing. As for the less redundant data such as RS and SP, all the statistical measures based on 'sequence space' achieve significantly improvement when word's length increases to 4 (Figures 4, 5 (k = 4)).

Question C

Using protein 'sequence space' contributes to the accuracy of protein classification. However, the construction of protein 'sequence space' relies heavily on the score matrix. In order to evaluate the influence of different score matrices, the function *MAUC(measure, score matrix)* is defined by

$$\begin{aligned}&\mathrm{MAUC}(\text{measure, score matrix})\\&\quad=\max_{1\le k\le 4}\left(\mathrm{AUC}(\text{measure, score matrix,k})\right),\end{aligned}\tag{2}$$

where *AUC (measure, score matrix, k)* denotes the area under ROC curve of the statistical measure based on the *k*-

Table 1: The entries of AUC for Data CK, RS and SP

CK		RS		SP	
Method	**Area**	**Method**	**Area**	**Method**	**Area**
NW-linear.b45	0.808	NW-linear.b40	0.605	NW-linear.b62	0.509
NW-affine.b45	**0.860**	NW-affine.b40	**0.614**	NW-affine.b40	0.540
SW-linear.b45	0.850	SW-linear.b40	0.600	SW-linear.b62	**0.548**
SW-affine.b45	0.850	SW-affine.b40	0.600	SW-affine.b62	**0.548**
Clustal X	0.807	Clustal X	0.555	Clustal X	0.535
k-word FPS[a]		***k-word FPS[a]***		***k-word FPS[a]***	
Method	**Area**	**Method**	**Area**	**Method**	**Area**
cos.1	0.729	cos.1	**0.609**	cos.1	0.569
eu.1	0.700	eu.1	0.607	eu.1	0.570
se.2	0.701	se.1	0.500	se.1	0.495
W.1.b45	0.652	W.1.b100	0.601	W.1.p120	0.559
s1.3	0.708	s1.2	0.581	s1.2	0.535
s2.3	0.708	s2.2	0.578	s2.3	0.530
gre.3	0.673	gre.1	0.607	gre.1	**0.572**
gsm.2	**0.791**	gsm.1	0.594	gsm.1	0.524
k-word FPSS[b]		***k-word FPSS[b]***		***k-word FPSS[b]***	
Method	**Area**	**Method**	**Area**	**Method**	**Area**
cos.3.p120	0.655	cos.1.p40	0.604	cos.1.p40	0.571
eu.1.p40	0.640	Eu.4.p80	0.603	eu.1.p40	0.570
se.3.b100	0.726	se.1.p250	0.501	se.2.p250	0.545
W.1.b45	0.652	W.1.b100	0.601	W.1.b100	0.559
s1.2.b40	0.667	s1.4.b40	0.607	s1.2.b100	0.554
s2.2.b40	0.667	s2.4.b40	0.607	s2.3.p40	0.545
gre.1.p40	0.683	Gre.4.b100	0.615	gre.1.p40	**0.575**
gsm.4.b100	**0.776**	gsm.3.b40	**0.627**	gsm.2.b100	0.557

The comparison of the areas under ROC curves (AUC) obtained from all the (dis)similarity measures for data CK, RS and SP. *k-word FPS[a]* denotes the *k*-word frequencies of protein sequences. *k-word FPSS[b]* denotes the *k*-word frequencies of protein 'sequence space'. All the abbreviations of (dis)similarity measures are illustrated in the "List of abbreviations" section.

word frequencies of protein 'sequence space' that is built based on the score matrix. The *MAUC* values of all the statistical measures based on ten score matrices for three data sets are presented in Figure 7. Figure 7 largely confirms that the measures possess different performances based on different score matrices. The changes of *DAUC* for the data CK, RS and SP are similar. For BLOSUM score matrix, BLOSUM40 and BLOSUM100 perform better in improvement of the statistical measures' classification abilities. As for PAM score matrix, PAM120 or PAM250 improves the classification ability of all the (dis)similarity measures on the high redundant data more obviously, except for the measures *eu.k* and *gre.k*. PAM40 or PAM80 contributes to improve the classification ability of the (dis)similarity measures more obviously on the less redundant data.

Phylogenetic analysis

Since *Gdis.k* is a statistical distance measure, it is further tested to analyze phylogenetic relationships. Given a set of protein sequences, their phylogenetic relationships can be obtained through the following main operations: firstly, the *k*-word frequencies of protein 'sequence space' are calculated; secondly, the statistical distances are calculated and arranged into a distance matrix; finally, the phylogenetic relationships is obtained by neighbor-joining program in the PHYLIP package [37].

A data set includes 68 SMC proteins, 5 Rad50 proteins and 5 MukB proteins (Additional file 4), which have been widely studied [38-42]. Our distance measure is applied to this data, and the results are shown in Figure 8. To assess the robustness of an estimated tree under perturbations of the input alignment, it is customary to perform a bootstrap analysis, where entire columns of the alignment

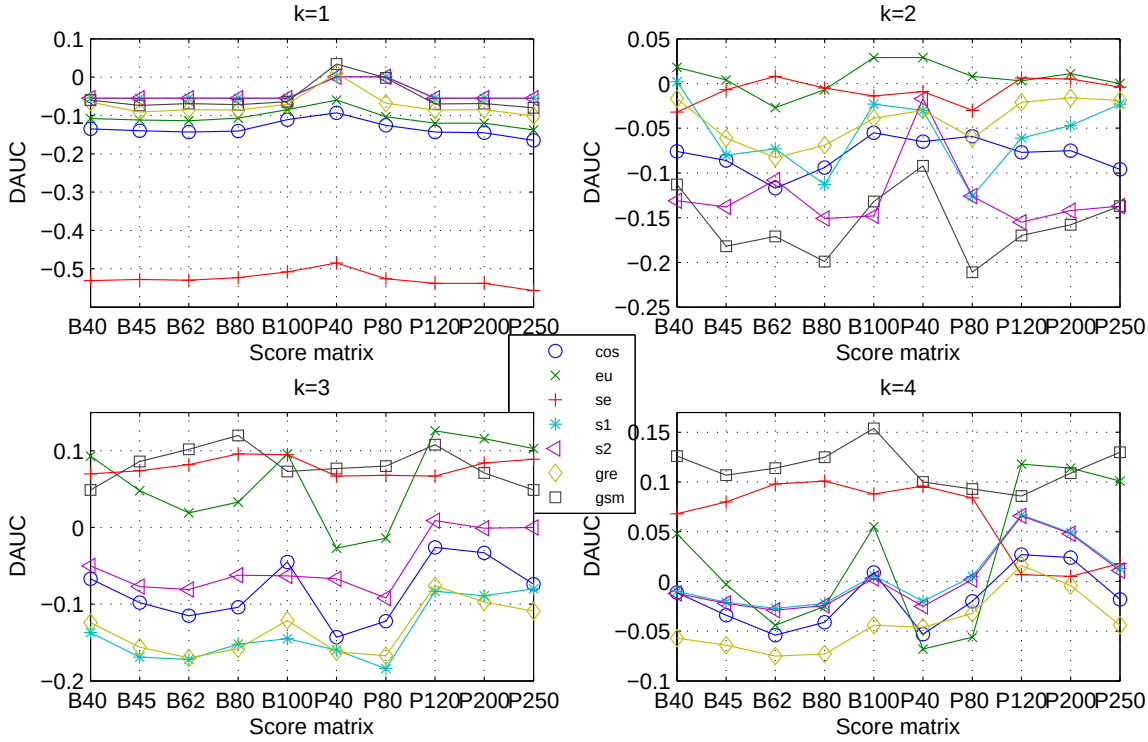

Figure 4
***DAUC* values for data CK**. The *DAUC* values of seven statistical measures for data CK. All statistical measures based on *k*-word frequencies of protein 'sequence space' run with *k* from 1 to 4, where protein 'sequence space' is constructed according to ten score matrices. One graph presents each word length (from 1 to 4).

are resampled with replacement. The bootstrap technique is employed to evaluate the tree topologies by resampling the sequence 100 times. We obtain the phylogenetic relationships drawn by MEGA program [11], bootstrap values, lower than 50, are hidden. Generally, an independent method can be developed to evaluate the accuracy of phylogenetic relationships, or the validity of phylogenetic relationships can be tested by comparing it with authoritative ones. Here, we adopt the latter one to test the validity of our measure.

Question D

Our results are quite consistent with the accepted taxonomy and authoritative ones [40-42] in the following three aspects. First of all, all the organisms are clearly separated from each other. Among the SMC proteins, it is consistently observed that SMC1 and SMC4 are grouped closely (there are the larger SMC subunits of the cohesin and condensin SMC heterodimers, respectively), and the smaller subunits, SMC3 and SMC2, appear to group closely. SMC5 and SMC6 are grouped together, which is consonant with that they heterodimerize as part of a DNA repair complex [42,43]. Secondly, it is obvious from this tree that the closest relatives to the SMC proteins are the Rad50 proteins, followed by MukB proteins. Many of these Rad50 superfamily proteins have the conserved N-terminal FKS (or FRS) motif (located before the Walker A site), which is presented in most of the SMC proteins [41]. Finally, among the SMC proteins, it is observed that SMC1 protein and SMC4 protein are closer to SMC proteins, followed by SMC2, SMC3, SMC5 and SMC6 [41,42]. It suggests that the duplication events giving rise to each subfamily must have occurred either before or very soon after the origin of eukaryotes. Since the rate of accepted amino acid substitution varies among different eukaryotic taxa within each subfamily. Condensin SMCs appear to show a higher substitution than cohesin SMCs, the mean distances within subfamilies of these proteins (averaged across all condensin and cohesin SMCs for each pairwise comparison between different organisms) are about half (0.54 ± 0.134) the corresponding distances between SMC5 and SMC6 proteins [41]. These reasonable results confirm that *Gdis.k* is a reliable distance measure for phylogenetic analysis.

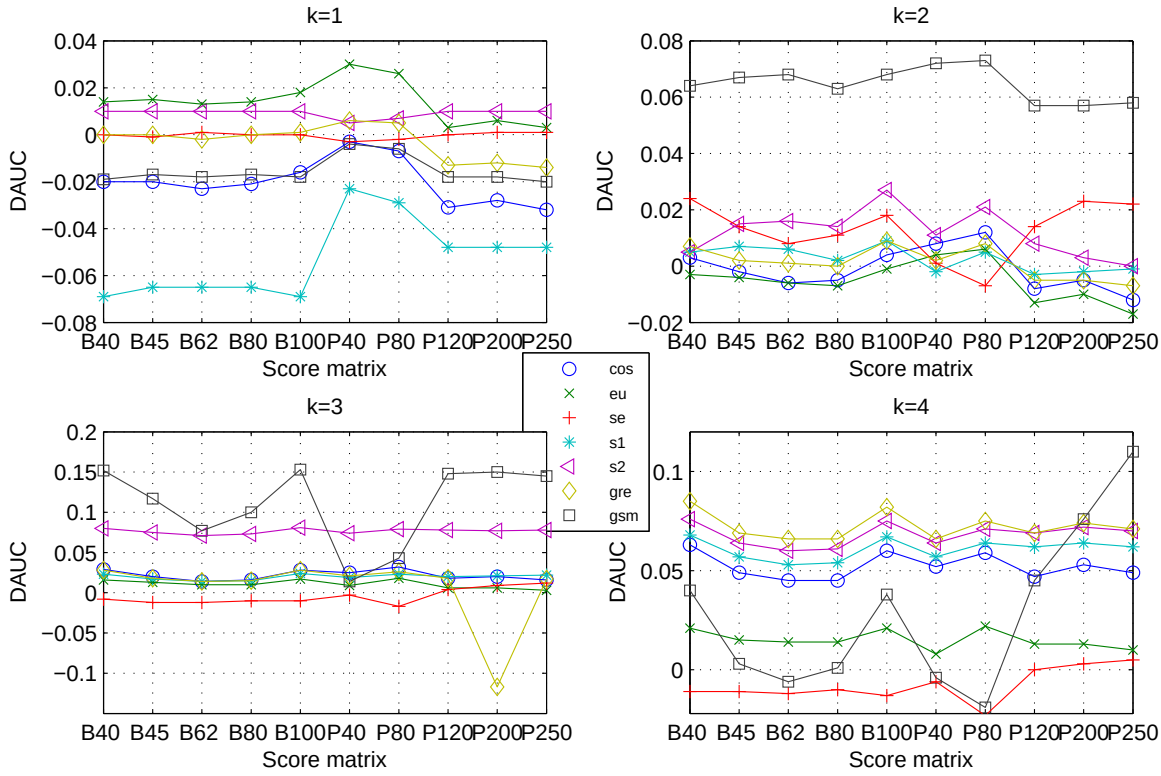

Figure 5
DAUC values for data RS. The *DAUC* values of seven statistical measures for data RS. All statistical measures based on *k*-word frequencies of protein 'sequence space' run with *k* from 1 to 4, where protein 'sequence space' is constructed according to ten score matrices. One graph presents each word length (from 1 to 4).

Conclusion

Prior to this research, the statistical measures are perceived as adequate for analysis of biological data mainly because of their flexibility and scalability with data set size. In particular, some of them are quantitatively compared for the recognition of SCOP relationships [14]. This article presents a novel way to compare protein sequences by exploring the information on 'sequence space' and two new statistical measures: *gre.k* and *gsm.k*. It offers the first systematic and quantitative experimental assessment of statistical measures based on protein sequence and protein 'sequence space', which naturally complements the many available comparisons based on protein sequences.

The accuracy of each (dis)similarity measure to classify protein sequence is assessed through the experiments on high redundant and less redundant data sets. The comparative index AUC is a good measure of overall accuracy of a classification scheme. The proposed statistical distance measure, *Gdis.k*, is further tested to analyze phylogenetic relationships.

As for the high redundant data, alignment-based measures have a clear advantage. *gsm.k*, followed by *cos.k*, is clearly more efficient among the existing statistical measures (Figure 1 and Table 1). When the data becomes less redundant, all the statistical measures, except for *se.k* and *s2.k*, outperform the alignment-based measures by exploring the information on protein 'sequence space', and *gre.k* proposed by us achieves the best performance (Figure 3 and Table 1). The scheme for constructing 'sequence space' can provide more information than the protein sequence only and contributes to the accuracy of protein classification, especially for the less redundant data sets such as RS and SP. Almost all the statistical measures based on 'sequence space' achieve significantly improvement when word's length increases to 4 (Figures 4, 5, 6). In addition, the reasonable results of phylogenetic analysis illustrate the validity of our distance measure for phylogenetic analysis.

Overall our comparison study highlights the necessity for alignment-free measures to extract more information as possible. Thus, this understanding can then be used to

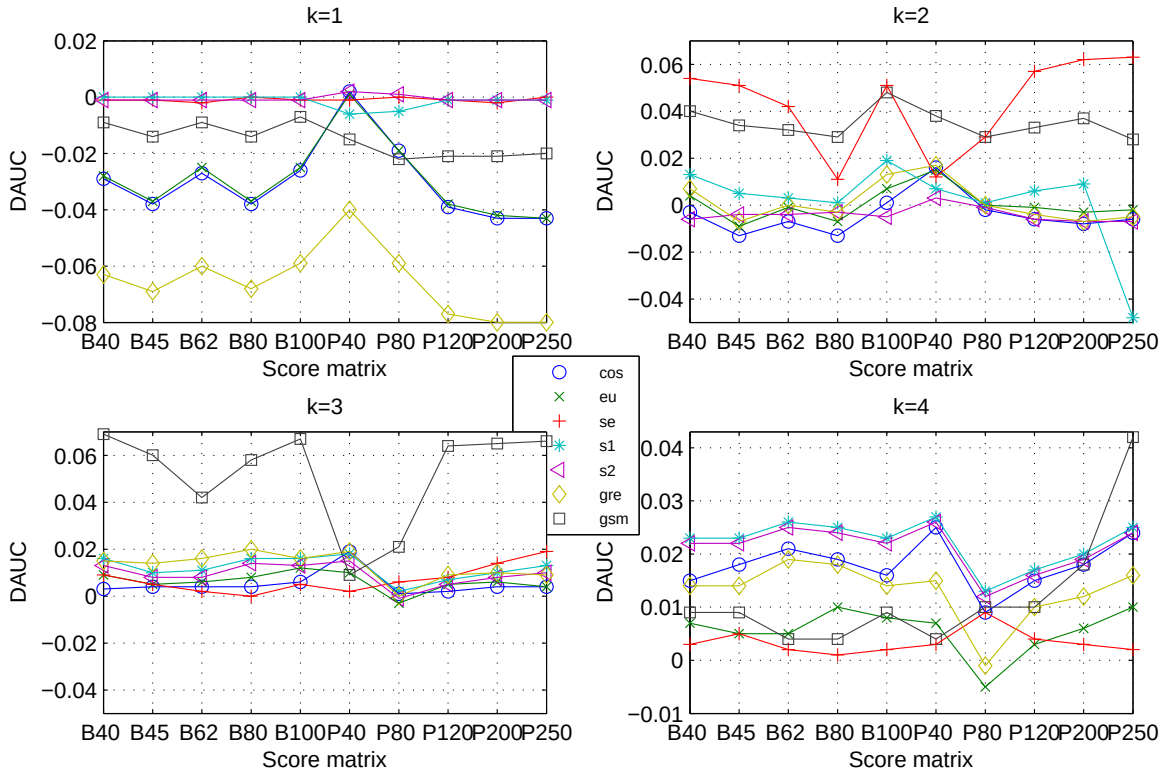

Figure 6
DAUC values for data SP. The *DAUC* values of seven statistical measures for data SP. All statistical measures based on *k*-word frequencies of protein 'sequence space' run with *k* from 1 to 4, where protein 'sequence space' is constructed according to ten score matrices. One graph presents each word length (from 1 to 4).

guide development of more powerful measures for protein sequence comparison with future possible improvement on evolutionary, structure and function study. But, it is worthy to note that although exploring the information on 'sequence space' improves the classification ability of some (dis)similarity measures, they all perform very poorly, near random classification values of 0.5 for less redundant data. That is to say, they may be useless in practice. So we expect a further investigation on the statistical methods, especially for low redundant datasets

Methods

Word statistics

Word statistics in protein sequence

There is a large body of literatures on word statistics [45], 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_k^s made up of *k*-word counts

$$C_k^s = \left(c(w_{k,1}), c(w_{k,1}), \cdots, c(w_{k,n}) \right), \qquad (3)$$

where *n* is the number of all possible *k*-words. For example, consider the protein sequence *s* = *VCST*, we can obtain the vector made up of *2*-word counts

$$C_2^s = \left(c(VC), c(CS), c(ST) \right) = (1,1,1). \qquad (4)$$

The frequencies of *k*-words, F_k^s, can he calculated by

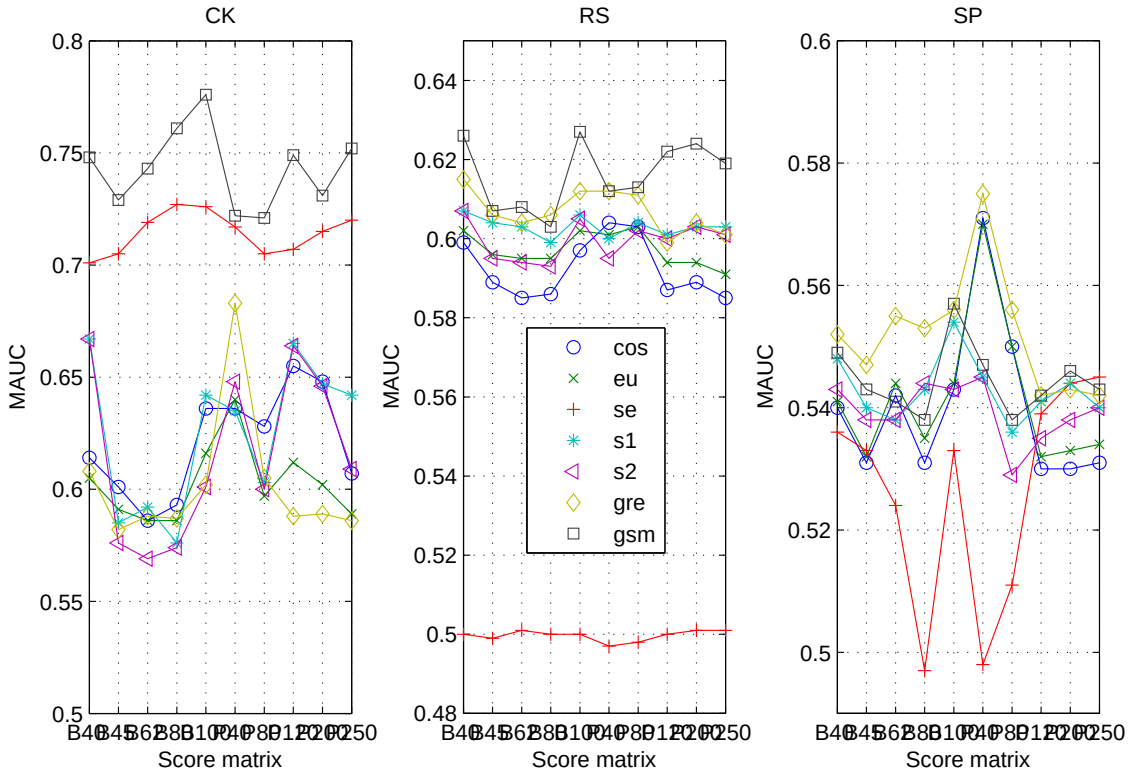

Figure 7
MAUC values for data sets CK, RS and SP. The *MAUC* values for the data CK, RS and SP, one for each data. All the statistical measures are based on *k*-word frequencies of protein 'sequence space', with ten score matrices to build protein 'sequence space'.

$$F_k^s = \left(f(w_{k,1}), f(w_{k,1}), \cdots, f(w_{k,n}) \right) = \left(\frac{c(w_{k,1})}{m-k+1}, \frac{c(w_{k,2})}{m-k+1}, \cdots, \frac{c(w_{k,n})}{m-k+1} \right). \quad (5)$$

Word statistics in protein 'sequence space'

The number of possible protein sequences is enormous. When a protein sequence is given, we are interested in its related proteins, and we denote them as the 'sequence space' of the given protein.

Substitution matrices represent similarity of amino acids, where each entry m_{ij} of a substitution matrix $[m_{ij}]$ represents the 'normalized probability' (score) that amino acid i can mutate into amino acid j. Let $i \aleph j$ denotes that the amino acids i and j are similar. Usually, two amino acids i and j are considered similar if $m_{ij} > 0$. That is to say

$$i \aleph j \text{ if } m_{ij} > 0 \ \forall \ i, j \in \Omega \quad (6)$$

where Ω = {A,C,D,E,F,G,H,I,K,L,M,N,P,Q,R,S,T,V,W,Y}. Note that the substitution matrices are symmetric matrices, i.e., a being similar to b implies that b is similar to a. But this similarity of amino acids is not a transitive relation. For example, a is similar to b and b is similar to c, but a is not similar to c. Therefore, 20 amino acids are not possibly classified into several similarity classes according to this property.

We shall bypass the above similarity classes and consider a new star set which is easily to implement. A star set assumes that the properties are known between vertices and center. We can construct a star set including all the vertices and the center, and specifically write the center as the first element of the set to distinguish one set from the others. For example, S is similar to A, T and N in BLOSUM62 substitution matrix, so S is the center and they can constitute a star set $\{S, A, T, N\}$ presented in Figure 9. For writing convenience, we write the star set $\{S, A, T, N\}$ as $\aleph S = \{x \mid x \aleph S, x \in \Omega\}$. With the aid of star set, 20 amino acids can be partitioned into 20 star sets pre-

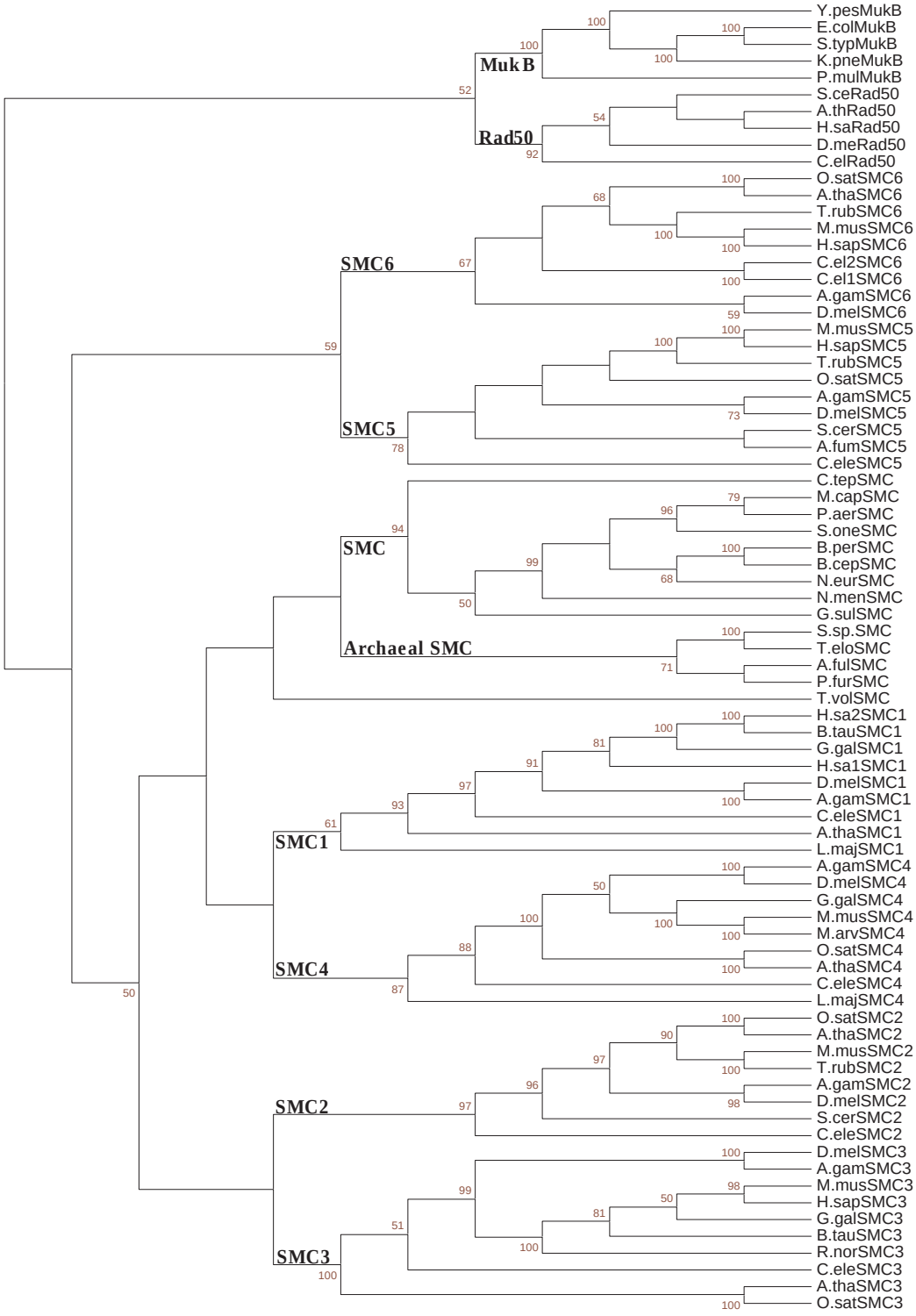

Figure 8

The diagram of phylogenetic relationships. Phylogenetic relationships are obtained by neighbor-joining program based on our statistical distance measure *Gdis.k* using all six SMC subfamilies, as well as the related MukB and Rad50. Bootstraps are based on 100 replications, and bootstrap values, lower than 50, are hidden.

sented in Table 2 based on BLOSUM62 substitution matrix.

Our work derives a way to build 'sequence space' with the help of star set. From the definition of star set, we know that each amino acid corresponds a star set. For example, the star set of the amino acid S is $\aleph S = \{S, A, T, N\}$ according to BLOSUM62 substitution matrix. Given two protein sequences $P = p_1p_2 \cup p_n$ and $Q = q_1q_2 \cup q_n$,

$$\forall\, p_i \in P,\, q_i \in Q, \text{ if } p_i \in \aleph q_i \Rightarrow P\ \Im\ Q \tag{7}$$

where $P\ \Im\ Q$ denotes that the protein sequences P and Q are related. Given a protein sequence s, its 'sequence space', denoted by SP_s, is defined as follows:

$$SP_s = \{P \mid P\ \Im\ s,\ length(P) = length(s)\} \tag{8}$$

where P is a protein sequence, $length(P)$ denotes the length of the protein sequence P. The protein 'sequence space' can be constructed as follows: for each protein sequence, beginning with the first amino acid, we scan through the protein sequence and substitute the star sets for amino acids at each position, respectively. Thus a special set of protein sequences is obtained, which is denoted as the 'sequence space' of the protein sequence. For example, given a protein sequence $s = VCST$, the star sets of V, C, S, and T are $\{V, M, I, L\}$, $\{C\}$, $\{S, A, T, N\}$ and $\{S, A, T, N\}$ according to BLOSUM62 substitution matrix, and the 'sequence space' of protein s is $\{V, M, I, L\}$-$\{C\}$-$\{S, A, T, N\}$-$\{T, S\}$.

Once the protein 'sequence space' is built, the k-word frequencies of 'sequence space' can be computed similarly. A segment of k symbols from a finite alphabet, A with 20 letters, is designated a k-word. The set $W_k = (w_{k,1}, w_{k,2}, \cup, w_{k,Y})$ consists of all possible k-words that can be extracted from protein 'sequence space', and has Y elements, where $Y = 20^k$. The count of k-words in protein 'sequence space', denoted by $C_k^{sp_s} = \left(c^{sp_s}(w_{k,1}), c^{sp_s}(w_{k,1}), \cdots, c^{sp_s}(w_{k,Y})\right)$ can be calculated by taking a sliding window with k-wide and scanning through the protein 'sequence space'. For example, considering the protein sequence $s = VCST$, its 'sequence space' is $\{V, M, I, L\}$-$\{C\}$-$\{S, A, T, N\}$-$\{T, S\}$, we can get a vector of 2-word counts

$$\begin{aligned} C_2^{SP_s} &= \left(c^{SP_s}(VC), c^{SP_s}(MC), \cdots, c^{SP_s}(NS)\right) \\ &= (1, 1, \cdots, 1) \end{aligned} \tag{9}$$

Similarly, one can then calculate k-word frequencies of protein 'sequence space', denoted as $F_k^{sp_s}$, by

$$\begin{aligned} F_k^{SP_s} &= \left(f^{SP_s}(w_{k,1}), f^{SP_s}(w_{k,1}), \cdots, f^{SP_s}(w_{k,n})\right) \\ &= \left(\frac{c^{SP_s}(w_{k,1})}{\sum_{t=1}^{Y} c^{SP_s}(w_{k,t})}, \frac{c^{SP_s}(w_{k,2})}{\sum_{t=1}^{Y} c^{SP_s}(w_{k,t})}, \cdots, \frac{c^{SP_s}(w_{k,Y})}{\sum_{t=1}^{Y} c^{SP_s}(w_{k,t})}\right). \end{aligned} \tag{10}$$

Table 2: The 20 star sets

Matrix	Star set				
BLOSUM62	{AS}	{C}	{DNE}	{EDQK}	{FYW}
	{G}	{HNY}	{IMLV}	{KEQR}	{LMIV}
	{MILV}	{NSDH}	{P}	{QERK}	{RQK}
	{SATN}	{T S}	{VMIL}	{WFY}	{YHFW}

With help of the star set, 20 amino acids are partitioned into 20 star sets based on substitution matrix BLOSUM62.

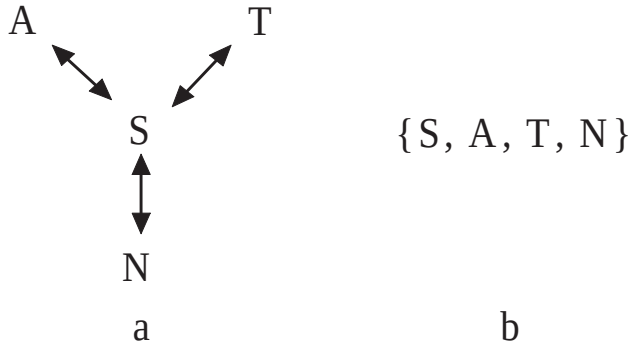

Figure 9
Representation of a star set. a: the diagram of star set, S is similar to A, T and N in BLOSUM62 substitution matrix, and S is the midpoint; b: the star set consists of the midpoint S and vertices A, T and N.

Statistical distance measures

Previous (dis)similarity measures

We first describe the six previous statistical measures for biological sequences.

Many statistical measures 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.

*1. Euclidian distance (*ed.k*)*

The Euclidian distance is one of the most common dissimilarity measures of biological sequences. The dissimilarity score between two protein sequences X and Y is the Euclidian distance between their k-word frequencies

$$F_k^A = \left(f(w_{k,1}^A), f(w_{k,1}^A), \cdots, f(w_{k,n}^A) \right) \quad \text{and}$$

$$F_k^B = \left(f(w_{k,1}^B), f(w_{k,1}^B), \cdots, f(w_{k,n}^B) \right) \text{[18]}$$

$$\begin{aligned} ed.k(X,Y) &= (F_k^X - F_k^Y)' \cdot (F_k^X - F_k^Y) \\ &= \sum_{t=1}^{n} (f(w_{k,t}^X) - f(w_{k,t}^Y))^2 . \end{aligned} \tag{11}$$

*2. Cosine of the angle (*cos.k*)*

In order to derive estimation of relatedness from the vector definitions of biological sequences, Stuart et al. (2002) proposed the pair-wise cosine for generating accurate gene and species phylogenies from whole genome sequences.

$$\cos .k(X,Y) = -\ln[(1+\cos(X,Y))/2],$$

$$\cos(X,Y) = \frac{\sum_{t=1}^{n} f(w_{k,t}^X) \cdot f(w_{k,t}^Y)}{\sqrt{\sum_{t=1}^{n} (f(w_{k,t}^X))^2} \cdot \sqrt{\sum_{t=1}^{n} (f(w_{k,t}^Y))^2}}. \tag{12}$$

Cosine is a standard measure of vector similarity, and its application for this purpose can be understood intuitively.

*3. Standardized Euclidean distance (*se.k*)*

The above measures explore the use of Euclidean distances and correlations between k-word frequencies representations of sequences. Standardized Euclidean distance takes into account the data covariance structure

$$\begin{aligned} se.k(X,Y) &= (F_k^X - F_k^Y)' \cdot [diag(s_{11}, \cdots s_{nn})]^{-1} (F_k^X - F_k^Y) \\ &= \sum_{t=1}^{n} \frac{f(w_{k,t}^X) - f(w_{k,t}^Y)}{s_{tt}} \end{aligned} , \tag{13}$$

where S = [s_{ij}] represents the covariance matrix of *k*-word frequencies. The standard Euclidean distance forces cov (f_i, f_j) = 0 for $i \neq j$. Therefore, in this distance measure the correlations between different *k*-words are ignored and only the same *k*-word variances are accounted for. The standard Euclidean distance was first proposed for sequence comparison by Wu et al. (1997).

*4. Kullback-Leibler discrepancy (*kld*)*

Let P_1 and P_2 be two probability frequencies on a universe *X*, the Kullback-Leibler divergence (*kld*) or the relative entropy, denoted as *kld(P1, P2)*, of P_1 with respect to P_2 is defined by the Lebesgue integral [46],

$$kld(P_1, P_2) = \int_X \log(\frac{d(P_1)}{d(P_2)}) d(P_1). \tag{14}$$

Although relative entropy is not a true metric, it satisfies many important mathematical properties. Wu et al. (2001) have applied Kullback-Leibler discrepancy to compare DNA sequences based on the frequencies of all *k*-words.

*5. W-metric (*W.k*)*

In an application where the covariance matrices *S* chosen in standard Euclidean distance is replaced by amino acid substitution matrices, Vinga et al. (2004) proposed and demonstrated the use of W-metric as a novel k-word composition metric

$$\begin{aligned} W.k(X,Y) &= (F_k^X - F_k^Y)' \cdot W \cdot (F_k^X - F_k^Y) \\ &= \sum_{i=1}^{n} \sum_{j=1}^{n} (f(w_{k,i}^X) - f(w_{k,i}^Y)) \cdot (f(w_{k,j}^X) - f(w_{k,j}^Y)) \cdot w_{ij} \end{aligned} \tag{15}$$

where *W* is amino acid substitution matrices such as BLOSUM and PAM. *W.k* is a distance defined between protein sequences, which bridges between alignment-based metrics and measures based solely on k-word composition.

6. S_1 *and* S_2 *(*s1.k *and* s2.k*)*

$S_1$and S_2 are statistical measures for protein sequences based on the concept of comparing the similarity between the *k*-word appearances [23]. If the set $W_k^X = \left(w_{k,1}^X, w_{k,2}^X, \cdots, w_{k,n}^X \right)$ and $W_k^Y = \left(w_{k,1}^Y, w_{k,2}^Y, \cdots, w_{k,n}^Y \right)$ consist of all possible *k*-words that can be extracted from proteins *X* and *Y*, respectively, S_1 and S_2 can be computed by

$$\begin{aligned} s1.k &= c \times |Match(W_k^X, W_k^Y)| \Big/ (|Word(W_k^X)| + |Word(W_k^Y)|) \\ s2.k &= c \times |Match(W_k^X, W_k^Y)| \Big/ (|Word(W_k^X)| + |Word(W_k^Y)| \\ &\quad - |length(X) - length(Y)|) \end{aligned} \tag{16}$$

where $|Match(W_k^X, W_k^Y)|$ is the total number of *k*-words shared by two proteins *X* and *Y*, constant c is a normalizing factor; $|Word(W_k^X)|$ and $|Word(W_k^Y)|$ denote the total numbers of occurred *k*-words in proteins *X* and *Y*.

Novel statistical distance measures

We describe two novel statistical measures for protein sequences comparison based on *k*-word frequencies.

*1. Generalized relative entropy (*gre.k*)*

Relative entropy is the most important concept in both statistical biology and information theory. It has been explored as similarity measures such as *kld* and *SimMM* [17,20] to compare biological sequences. However, in an application where P_k is equal to 0 or 1, $kld(P^1, P^2) \to \infty$. So the similarity measure *kld* becomes unsuitable. For such an application, we generalize relative entropy with the help of Jensen-Shannon Divergence, denoted by *gre.k*, by

$$gre.k(X,Y) = \sum_{t=1}^{n} f(w_{k,t}^X) \cdot \log_2 \left(\frac{2 \cdot f(w_{k,t}^X)}{f(w_{k,t}^X) + f(w_{k,t}^Y)} \right) \tag{17}$$

Now, if $f(w_{k,t}^X)$ is equal to 0 and 1,

$$f(w_{k,t}^X) \cdot \log_2 \left(\frac{2 \cdot f(w_{k,t}^X)}{f(w_{k,t}^X) + f(w_{k,t}^Y)} \right) = 0. \tag{18}$$

So *gre.k* can deal with all kinds of *k*-word frequencies.

*2. Gapped similarity measure (*gsm.k*)*

From the definition of *gre.k*, it is worthy to note that the frequencies of *k*-words that are present in both sequences have different impact on the *gre.k*. But the frequencies of *k*-words that are present in only one sequence have no contribution to *gre.k*. Because if $f(w_{k,t}^X)$ or $f(w_{k,t}^Y)$ is equal to 0,

$$f(w_{k,t}^X) \cdot \log_2 \left(\frac{2 \cdot f(w_{k,t}^X)}{f(w_{k,t}^X) + f(w_{k,t}^Y)} \right) = 0 \quad \text{or} \quad f(w_{k,t}^Y) \cdot \log_2 \left(\frac{2 \cdot f(w_{k,t}^Y)}{f(w_{k,t}^X) + f(w_{k,t}^Y)} \right) = 0.$$

Similarly, the measures S_1 and S_2 focus on the appearances of k-words but ignore their frequencies. Motivated by extracting the information from all the *k*-words, we investigate a novel statistical measure for protein sequence comparison, called the gapped similarity measure

$$gsm.k(X,Y) = \sum_{t=1}^{n} score(f(w_{k,t}^X), f(w_{k,t}^Y))$$

$$score(f(w_{k,t}^X), f(w_{k,t}^Y)) = \begin{cases} f(w_{k,t}^X) \cdot \log_2 \left(\frac{2 \cdot f(w_{k,t}^X)}{f(w_{k,t}^X) + f(w_{k,t}^Y)} \right) & \text{if } f(w_{k,t}^X) \neq 0 \text{ and } f(w_{k,t}^Y) \neq 0 \\ 1 & \text{if } f(w_{k,t}^X) = 0 \text{ and } f(w_{k,t}^Y) \neq 0 \\ 1 & \text{if } f(w_{k,t}^X) \neq 0 \text{ and } f(w_{k,t}^Y) = 0 \\ 0 & \text{if } f(w_{k,t}^X) = 0 \text{ and } f(w_{k,t}^Y) = 0 \end{cases} \tag{19}$$

In the definition of function *score*, the frequencies of all the *k*-words in protein sequence are considered. Indeed, the measure *gsm.k* is the edit score between *k*-word frequencies of the two protein sequences *X* and *Y*. If a *k*-word *w* appears in the two sequences, the edit score is $f(w_{k,w}^X) \cdot \log_2 \left(\frac{2 \cdot f(w_{k,w}^X)}{f(w_{k,w}^X) + f(w_{k,w}^Y)} \right)$. If a *k*-word *w* appears in protein sequence *X* not *Y*, it seems that the *k*-word *w* is deleted from the protein sequence *Y*, we choose the maximum value of function

$f(w_{k,w}^X) \cdot \log_2 \left(\frac{2 \cdot f(w_{k,w}^X)}{f(w_{k,w}^X) + f(w_{k,w}^Y)} \right)$ as the gap penalty according to followed proposition.

Proposition. If $F_k^A = \left(f(w_{k,1}^A), f(w_{k,1}^A), \cdots, f(w_{k,n}^A) \right)$ *and* $F_k^B = \left(f(w_{k,1}^B), f(w_{k,1}^B), \cdots, f(w_{k,n}^B) \right)$ are two *k*-word frequency vectors of length *n*,

$$f(w_{k,t}^X) \cdot \log_2 \left(\frac{2 \cdot f(w_{k,t}^X)}{f(w_{k,t}^X) + f(w_{k,t}^Y)} \right) \leq 1. \tag{20}$$

Proof: To find its maximum, we rewrite

$$f(w_{k,t}^X) \cdot \log_2 \left(\frac{2 \cdot f(w_{k,t}^X)}{f(w_{k,t}^X) + f(w_{k,t}^Y)} \right) = f(w_{k,t}^X) \cdot \log_2 2 + f(w_{k,t}^X) \cdot \log_2 \left(\frac{f(w_{k,t}^X)}{f(w_{k,t}^X) + f(w_{k,t}^Y)} \right).$$

Since $\frac{f(w_{k,t}^X)}{f(w_{k,t}^X) + f(w_{k,t}^Y)} \leq 1$, we can get

$$\log_2 \left(\frac{f(w_{k,t}^X)}{f(w_{k,t}^X) + f(w_{k,t}^Y)} \right) \leq 0.$$

Thus

$$f(w_{k,t}^X) \cdot \log_2 \left(\frac{2 \cdot f(w_{k,t}^X)}{f(w_{k,t}^X) + f(w_{k,t}^Y)} \right) \leq 1.$$

Similarly, the symmetric form of *gsm.k*, denoted as *Gdis.k*, between two sequences *X* and *Y* is defined by

$$Gdis.k(X,Y) = \begin{cases} 0 & if\ F_k^X = F_k^Y \\ (gsm(X,Y) + gsm(Y,X))/n + 2 & else \end{cases}. \quad (21)$$

A distance metric, $D(\cdot,\cdot)$, should satisfy the following conditions:

1. $D(S, Q) \geq 0$, where the equality is satisfied iff $S = Q$ (identity).

2. $D(S, Q) = D(Q, S)$(symmetry).

3. $D(S, Q) \leq D(S, T) + D(T, Q)$(triangle inequality).

In the appendix, we prove that the statistical measure, *Gdis.k*, defined above satisfies the three conditions and is, therefore, a valid distance metric.

Evaluation methods

Similarity/dissimilarity measures are compared by considering how well they classify protein sequences, as well as by computing receiver operator characteristic (ROC) curves. ROC goes back to signal detection and classification problems and is now widely used [47]. 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. 22].

$$sensitivity = \frac{TruePositives}{Positives} = \frac{TP}{TP+FN},$$
$$specificity = \frac{TrueNegatives}{Negatives} = \frac{TN}{TN+FP}, \quad (22)$$
$$1 - specificity = \frac{FP}{TN+FP}.$$

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 classificator because it is a threshold independent performance measure and is closely related to the Wilcoxon signed-rank test [48]. For a perfect classifier, the AUC is 1 and for a random classifier the AUC is 0.5

Availability

Software name: SMPS-SS

Software home page: http://math.dlut.edu.cn/daiqi/SMPS-SS.html

Operating system(s): windows

Programming languages: perl

License: web server freely available without registration

Restrictions to use by non-academics: on request

Abbreviations

AUC: Area Under the Curve; b..: BLOSUM..; CATH: Hierarchical Classification of Protein Domain Structures; CD-HIT: Cluster Database at High Identity with Tolerance; CK: Chew-Kedem Data; cos.*k*: Cosine of the Angle Based on *k*-word Frequencies of Protein Sequence; cos,*k*.matrix: Cosine of the Angle Based on *k*-word Frequencies of Protein 'Sequence Space' Constructed According to Score Matrix; DAUC: Different Area under the Curve; ed.*k*: Euclidian Distance Based on *k*-word Frequencies of Protein Sequence; ed.*k*.matrix: Euclidian Distance Based on *k*-word Frequencies of Protein 'Sequence Space' Constructed According to Score Matrix; FP: False Positives; Gdis.*k*: Gapped Distance Measure Based on *k*-word Frequencies; gre.*k*: Generalized Relative Entropy Based on *k*-word Frequencies of Protein Sequence; gre.*k*.matrix: Generalized Relative Entropy Based on *k*-word Frequencies of Protein 'Sequence Space' Constructed According to Score Matrix; gsm.*k*: Gapped Similarity Measure Based on *k*-word Frequencies of Protein Sequence; gsm.*k*.matrix: Gapped Similarity Measure Based on *k*-word Frequencies of Protein 'Sequence Space' Constructed According to Score Matrix; kld: Kullback-Leibler Discrepancy; MAUC: Maximal Area under the Curve; MEGA: Molecular Evolutionary Genetics Analysis; NW: Needleman-Wunsch Measure; NW-linear.matrix: Needleman-Wunsch Measure Using Score Matrix and Linear Gap Penalty; NW-affine.matrix: Needleman-Wunsch Measure Using Score Matrix and Affine Gap Penalty; p..: PAM..; pfam: Protein Family; PIR: Protein Information Resource; ROC: Receiver Operating Curve; RS: Rost and Sander Data; s1.k: S1 Measure Based on *k*-word Frequencies of Protein Sequence; s1.matrix: S1 Measure Based on *k*-word Frequencies of Protein 'Sequence Space' Constructed According to Score Matrix; s2.k: S2 Measure Based on *k*-word Frequencies of Protein Sequence; s2.k.matrix: S2 Measure Based on *k*-word Frequencies of Protein 'Sequence Space' Constructed According to Score Matrix; SCOP: Structural Classification of Proteins; se.k: Standardized Euclidean Distance Based on *k*-word Frequencies of Protein Sequence; se.k.matrix: Standardized Euclidean Distance Based on *k*-word Frequencies of Protein 'Sequence Space' Constructed According to Score Matrix; SM: Smith-Waterman Measure; SM-linear.matrix: Smith-Waterman Meas-

ure Using Score Matrix and Linear Gap Penalty; SM-affine.matrix: Smith-Waterman Measure Using Score Matrix and Affine Gap Penalty; SMC: Structural Maintenance of Chromosomes; SP: Sierk-Pearson Data; SPs: 'Sequence Space' of Sequence *s*; SS.k: Similarity Score Based on *k*-word Frequencies; Swiss-Prot: Swiss-Prot Database; TP: True Positives; W.k.matrix: W-metric Based on *k*-word Frequencies and Score Matrix;

Authors' contributions

QD conceived the method and prepared the manuscript. QD implemented the software and performed the ROC analysis. QD and TMW contributed to the discussion and have approved the final manuscript.

Appendix

The proof of valid distance metric

Lemma 1. For a real convex function f in its domain [a, b], $\forall\, x_i \in [a, b]$, $\lambda_i > 0$ (i = 1, 2, ∪, n), $\sum_{i=1}^{n} \lambda_i = 1$, Jensen's inequality can be stated as:

$$f(\sum_{i=1}^{n} \lambda_i x_i) \le \sum_{i=1}^{n} \lambda_i f(x_i). \tag{23}$$

Proof: Let $x_0 = \sum_{i=1}^{n} \lambda_i x_i$, $x_0 \in [s, b]$. We expand f(x) around x_0, and by Taylor's theorem, we have that

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{f''(\xi)}{2!}(x - x_0)^2, \quad \xi \in [a, b].$$

Since f(x) is a real convex function f in its domain [a, b], $f''(\xi) > 0$. Thus we have

$$f(x) \ge f(x_0) + f'(x_0)(x - x_0).$$

For all $x_i \in [a, b]$, we can obtain that

$$\begin{aligned} f(x_1) &\ge f(x_0) + f'(x_0)(x_1 - x_0), \\ f(x_2) &\ge f(x_0) + f'(x_0)(x_2 - x_0), \\ &\vdots \\ f(x_n) &\ge f(x_0) + f'(x_0)(x_n - x_0). \end{aligned}$$

Multiplying the above inequalities with λ_i, we have

$$\begin{aligned} \lambda_1 \cdot f(x_1) &\ge \lambda_1 \cdot f(x_0) + \lambda_1 \cdot f'(x_0)(x_1 - x_0), \\ \lambda_2 \cdot f(x_2) &\ge \lambda_2 \cdot f(x_0) + \lambda_2 \cdot f'(x_0)(x_2 - x_0), \\ &\vdots \\ \lambda_n \cdot f(x_n) &\ge \lambda_n \cdot f(x_0) + \lambda_n \cdot f'(x_0)(x_n - x_0). \end{aligned}$$

Summing the above inequalities,

$$\begin{aligned} \sum_{i=1}^{n} \lambda_i f(x_i) &\ge f(x_0) \sum_{i=1}^{n} \lambda_i + \sum_{i=1}^{n} \lambda_i f'(x_0)(x_i - x_0) \\ &= f(x_0). \end{aligned}$$

Thus, we obtain that

$$f(\sum_{i=1}^{n} \lambda_i x_i) \le \sum_{i=1}^{n} \lambda_i f(x_i).$$

Proposition 1. $\forall\, x, y > 0$,

$$(x + y) \ln\left(\frac{x+Y}{2}\right) \le x \ln x + y \ln y. \tag{24}$$

Proof: Let $f(x) = x\ln x$, $x > 0$, we calculate $f'(x)$ and $f''(x)$,

$$f'(x) = \ln x = 1, \quad f''(x) = \frac{1}{x}.$$

Thus $f(x)$ is a real convex function.

According to Lemma 1, we have

$$\frac{x+y}{2} \ln\left(\frac{x+Y}{2}\right) \le \frac{1}{2}(x \ln x + y \ln y).$$

Then

$$(x + y) \ln\left(\frac{x+Y}{2}\right) \le (x \ln x + y \ln y).$$

If $F_k^A = \left(f(w_{k,1}^A), f(w_{k,1}^A), \cdots, f(w_{k,n}^A)\right)$ *and* $F_k^B = \left(f(w_{k,1}^B), f(w_{k,1}^B), \cdots, f(w_{k,n}^B)\right)$ are two *k*-word frequency vectors of protein sequences *X* and *Y*, respectively, we define similarity score, denoted by *ss.k*, as follows:

$$\begin{aligned} ss.k(X, Y) = &\sum_{t=1}^{n} score(f(w_{k,t}^X), f(w_{k,t}^Y)) \\ &+ \sum_{t=1}^{n} score(f(w_{k,t}^Y), f(w_{k,t}^X)) \end{aligned}, \tag{25}$$

where

$$score(f(w_{k,t}^I), f(w_{k,t}^J)) = \begin{cases} f(w_{k,t}^I) \cdot \log_2\left(\frac{2 \cdot f(w_{k,t}^I)}{f(w_{k,t}^I) + f(w_{k,t}^J)}\right) & \text{if } f(w_{k,t}^I) \ne 0 \text{ and } f(w_{k,t}^J) \ne 0 \\ 1 & \text{if } f(w_{k,t}^I) = 0 \text{ and } f(w_{k,t}^J) \ne 0 \\ 1 & \text{if } f(w_{k,t}^I) \ne 0 \text{ and } f(w_{k,t}^J) = 0 \\ 0 & \text{if } f(w_{k,t}^I) = 0 \text{ and } f(w_{k,t}^J) = 0 \end{cases}$$

Proposition 2.

$$0 \le score(f(w_{k,t}^I), f(w_{k,t}^J)) + f(w_{k,t}^I), f(w_{k,t}^J) \le 2. \quad (26)$$

Proof: Firstly, we need to show that

$$score(f(w_{k,t}^I), f(w_{k,t}^J)) + f(w_{k,t}^J), f(w_{k,t}^I) \ge 0. \quad (27)$$

Case 1: $f(w_{k,t}^I) = f(w_{k,t}^J) = 0$, it satisfies the above inequality.

Case 2: The entry of $f(w_{k,t}^I)$ or $f(w_{k,t}^J)$ is equal to zero. Without loss of generality, assume $f(w_{k,t}^I) = 0$ and $f(w_{k,t}^J) \neq 0$, we can easily get that

$$score(f(w_{k,t}^I), f(w_{k,t}^J)) + f(w_{k,t}^J), f(w_{k,t}^I) = 2 > 0.$$

Case 3: $f(w_{k,t}^I) \neq 0$ and $f(w_{k,t}^J) \neq 0$. Using the Proposition 1, we can easily obtain the inequality (24).

To find its maximum, we use the Proposition in Method section to get that

$$score(f(w_{k,t}^I), f(w_{k,t}^J)) + f(w_{k,t}^I), f(w_{k,t}^J) \le 2.$$

Theorem 1. The statistical measure *Gdis.k(X,Y)* is a distance metric.

Proof: Again, by definition $ss.k(X, Y)$ and Proposition 2, we can obtain that it satisfies two important mathematical properties: (1) positivity: $Gdis.k(X, Y) \ge 0$ and $Gdis.k(X, Y) = 0 \Leftrightarrow F_k^X = F_k^Y$; (2) symmetry: $Gdis.k(X, Y) = Gdis.k(Y, X)$. We now need to show that $Gdis.k(X, Y) \ge 0$ satisfies the triangle inequality:

$$Gdis.k(X, Y) \le Gdis.k(X, Z) + Gdis.k(Z, Y).$$

Case 1: $F_k^X = F_k^Y = F_k^Z$, it satisfies the triangle inequality.

Case 2: Among three *k*-word frequency vectors, two vectors are equal. Without loss of generality, assume $F_k^X \neq F_k^Y$ and $F_k^X = F_k^Z$, we can easily obtain that

$$Gdis.k(X, Y) \le Gdis.k(X, Z) + Gdis.k(Z, Y).$$

Case 3: $F_k^X \neq F_k^Y \neq F_k^Z$. From the definition of *ss.k* and Proposition 2, we have

$$Gdis.k(X, Z) + Gdis.k(Z, Y) = ss.k(X, Z)/n + ss.k(Z, Y)/n + 4 \ge 4.$$

Since

$$Gdis.k(X, Y) = ss.k(X, Y)/n + 2 \le 4.$$

Thus

$$Gdis.k(X, Y) \le Gdis.k(X, Z) + Gdis.k(Z, Y).$$

Additional material

Additional file 1

The Chew-Kedem data set. *The protein sequences used in Chew-Kedem data with the accession numbers of PDB.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-9-394-S1.pdf]

Additional file 2

The Rost-Sander dataset. *The protein sequences used in Rost-Sander data with the accession numbers of PDB.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-9-394-S2.pdf]

Additional file 3

The Sierk-Pearson data. *The protein sequences used in Sierk-Pearson data with the accession numbers of PDB.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-9-394-S3.pdf]

Additional file 4

The protein data used in phylogenetic analysis. *The protein sequences used in phylogenetic analysis with abbreviated names, full names and Accession numbers.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-9-394-S4.pdf]

Acknowledgements

The authors thank all the anonymous referees for their valuable suggestions and support. In particular, the authors thank Prof. Susana Vinga for providing the MATLAB code for W-metric.

References

1. Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, Griffiths JS, Howe KL, Marshall M, Sonnhammer ELL: **The Pfam Protein FamiliesDatabase.** *Nucleic Acids Res* 2002, **30:**276-280.
2. Andreeva A, Howorth D, Brenner SE, Hubbard TJP, Chothia C, Murzin AG: **SCOP database in refinements integrate structure and sequence family data.** *Nucleic Acid Res* 2004, **32:**D226-D229.

3. Bairoch A, Apweiler R: **The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000.** *Nucleic Acids Res* 2000, **28:**45-48.
4. Wu CH, Huang H, Arminski L, Castro-Alvear J, Chen Y, Hu ZZ, Ledley RS, Lewis KG, Mewes HW, Orcutt BC, Suzek BE, Tsugita A, Vinayaka CR, Yeh LSL, Zhang J, Barker WC: **The Protein Information Resource, an integrated public resource of functional annotation of proteins.** *Nucleic Acids Res* 2002, **30:**35-37.
5. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215:**403-410.
6. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25:**3389-3402.
7. Pham TD: **Spectral distortion measures for biological sequence comparisons and database searching.** *Pattern Recog* 2007, **40:**516-529.
8. Felsenstein J: **Evolutionary trees from DNA sequences, a maximum likelihood approach.** *J Mol Evol* 1981, **17:**368-376.
9. Felsenstein J: **Inferring phylogenies from protein sequences by parsimony, distance and likelihood methods.** *Meth Enzymol* 1996, **266:**418-427.
10. Huelsenbeck JP, Ronquist F: **MRBAYES: Bayesian inference of phylogenetic trees.** *Bioinformatics* 2001, **17:**754-755.
11. Kumar S, Tamura K, Nei M: **MEGA3: integrated software for molecular evolutionary genetics analysis and sequence alignment.** *Brief Bioinform* 2004, **5(2):**150-163.
12. Ronquist F, Huelsenbeck JP: **MrBayes 3: Bayesian phylogenetic inference under mixed models.** *Bioinformatics* 2003, **19:**1572-1574.
13. Komatsu K, Zhu S, Fushimi H, Qui TK, Cai S, Kadota S: **Phylogenetic analysis based on 18S rRNA gene and matK gene sequences of Panax vietnamensis and five related species.** *Planta Med* 2001, **67:**461-465.
14. Vinga S, Gouveia-Oliveira R, Almeida JS: **Comparative evaluation of word composition distances for the recognition of SCOP relationships.** *Bioinformatics* 2004, **20(2):**206-15.
15. Ferragina P, Giancarlo R, Greco V, Manzini G, Valiente G: **Compression-based classification of biological sequences and structures via the Universal Similarity Metric: experimental assessment.** *BMC Bioinformatics* 2007, **8:**252-272.
16. Vinga S, Almeida J: **Alignment-free sequence comparison – a review.** *Bioinformatics* 2003, **19:**513-523.
17. Pham TD, Zuegg J: **A probabilistic measure for alignment-free sequence comparison.** *Bioinformatics* 2004, **20:**3455-3461.
18. Blaisdell BE: **A measure of the similarity of sets of sequences not requiring sequence alignmen.** *Proc Natl Acad Sci USA* 1986, **83:**5155-5159.
19. Wu TJ, Burke JP, Davison DB: **A measure of DNA sequence dissimilarity based on Mahalanobis distance between frequencies of words.** *Biometrics* 1997, **53:**1431-1439.
20. Wu TJ, Hsieh YC, Li LA: **Statistical measures of DNA dissimilarity under Markov chain models of base composition.** *Biometrics* 2001, **57:**441-448.
21. Stuart GW, Moffett K, Baker S: **Integrated gene and species phylogenies from unaligned whole genome protein sequences.** *Bioinformatics* 2002, **18:**100-108.
22. Fichant G, Gautier C: **Statistical method for predicting protein coding regions in nucleic acid sequences.** *Comput Appl Biosci* 1987, **3:**287-295.
23. Wu KP, Lin HN, Sung TY, Hsu WL: **A New Similarity Measure among Protein Sequences.** *Proceedings of IEEE CSB2003 Computer Society Bioinformatics Conference* 2003:347-352.
24. Didier G, Laprevotte I, Pupin M, Hénaut A: **Local decoding of sequences and alignment-free comparison.** *J Comput Biol* 2006, **13:**1465-1476.
25. Kelil A, Wang S, Brzezinski R, Fleury A: **CLUSS: Clustering of Protein Sequences Based on a New Similarity Measure.** *BMC Bioinformatics* 2007, **8:**286-305.
26. Hochreiter S, Heusel M, Obermayer K: **Fast model-based protein homology detection without alignment.** *Bioinformatics* 2007, **23:**1728-1736.
27. Chew LP, Kedem K: **Finding the Consensus Shape for a Protein Family.** *Algorithmica* 2003, **38:**115-129.
28. Sierk M, Person W: **Sensitivity and Selectivity in Protein Structure Comparison.** *Protein Sci* 2004, **13(3):**773-785.
29. Thiruv B, Quon G, Saldanha SA, Steipe B: **Nh3D: A Reference Dataset of Non-Homologous Protein Structures.** *BMC Struct Biol* 2005, **5:**12.
30. Word JM, Lovell SC, LaBean TH, Taylor HC, Zalis ME, Presley BK, Richardson JS, Richardson DC: **Visualizing and Quantifying Molecular Goodness-of-Fit: Small-Probe Contact Dots with Explicit Hydrogen Atoms.** *J Mol Biol* 1999, **285(4):**1711-1733.
31. Krasnogor N, Pelta DA: **Measuring the Similarity of Protein Structures by Means of the Universal Similarity Metric.** *Bioinformatics* 2004, **20(7):**1015-1021.
32. Rost B, Sander C: **Prediction of protein secondary structure at better than 70% accuracy.** *J Mol Biol* 1993, **232:**584-599.
33. Barthel D, Hirst JD, Błażewicz J, Burke EK, Krasnogor N: **ProCKSI: A Decision Support System for Protein (Structure) Comparison, Knowledge, Similarity and Information.** *BMC Bioinformatics* 2007, **8:**416.
34. **SCOP: Structural Classification of Proteins** [http://scop.mrclmb.cam.ac.uk/scop]
35. Pearl F, *et al.*: **The CATH Domain Structure Database and Related Resources Gene3D and DHS Provide Comprehensive Domain Family Information for Genome Analysis.** *Nucleic Acids Res* 2005, **33(D):**D247-D251.
36. Li W, Godzik A, Cd-hit: **A fast program for clustering and comparing large sets of protein or nucleotide sequences.** *Bioinformatics* 2006, **22:**1658-1659.
37. Felsenstein J: **PHYLIP-Phylogeny inference package (version 3.2).** *Cladistics* 1989, **5:**164-166.
38. Saitoh N, Goldberg I, Earnshaw WC: **The SMC proteins and the coming of age of the chromosome scaffold hypothesis.** *BioEssays* 1995, **17:**759-766.
39. Lowe J, Cordell SC, Ent F Van Den: **Crystal structure of the SMC head domain: an ABC ATPase with 900 residues antiparallel coiled-coil inserted.** *J Mol Biol* 2001, **306:**25-35.
40. Hirano M, Hirano T: **Hinge-mediated dimerization of SMC protein is essential for its dynamic interaction with DNA.** *EMBO J* 2002, **21:**5733-5744.
41. Cobbe N, Heck MM: **SMCs in the world of chromosome biology- from prokaryotes to higher eukaryotes.** *J Struct Biol* 2000, **129:**123-143.
42. Soppa J: **Prokaryotic structural maintenance of chromosomes (SMC) proteins: distribution, phylogeny, and comparison with MukBs and additional prokaryotic and eukaryotic coiled-coil proteins.** *Gene* 2001, **278:**253-264.
43. Taylor EM, Moghraby JS, Lees JH, Smit B, Moens PB, Lehmann AR: **Characterization of a novel human SMC heterodimer homologous to the Schizosaccharomyces pombe Rad18/Spr18 complex.** *Mol Biol Cell* 2001, **12:**1583-1594.
44. Fujioka Y, Kimata Y, Nomaguchi K, Watanabe K, Kohno K: **Identification of a novel non-SMC component of the SMC5/SMC6 complex involved in DNA repair.** *J Biol Chem* 2002, **277:**21585-21591.
45. Reinert G, Schbath S, Waterman MS: **Probabilistic and statistical properties of words: an overview.** *J Comput Biol* 2000, **7:**1-46.
46. Kroupa T: **Measure of divergence of possibility measures.** *Proceedings of the 6th Workshop on Uncertainty Processing (WUPES'2003), Hejnice, Czech Republic* :173-181.
47. Egan JP: *Signal Detection Theory and ROC-Analysis* Academic Press, New York; 1975.
48. Bradley AP: **The use of the area under the ROC curve in the evaluation of machine learning algorithms.** *Pattern Recog* 1997, **30:**1145-1159.