RESEARCH ARTICLE

# An Alignment-Free Algorithm in Comparing the Similarity of Protein Sequences Based on Pseudo-Markov Transition Probabilities among Amino Acids

**Yushuang Li[1], Tian Song[1]*, Jiasheng Yang[2], Yi Zhang[3], Jialiang Yang[4]***

**1** School of Science, Yanshan University, Qinhuangdao, China, **2** Department of Civil and Environmental Engineering, National Universality of Singapore, Singapore, **3** Department of Mathematics, Hebei University of Science and Technology, Shijiazhuang, Hebei, China, **4** School of Mathematics and Information Science, Henan Polytechnic University, Henan, China

* jialiang.yang@mssm.edu (JLY); songtian2012@126.com (TS)

OPEN ACCESS

**Citation:** Li Y, Song T, Yang J, Zhang Y, Yang J (2016) An Alignment-Free Algorithm in Comparing the Similarity of Protein Sequences Based on Pseudo-Markov Transition Probabilities among Amino Acids. PLoS ONE 11(12): e0167430. doi:10.1371/journal.pone.0167430

**Editor:** Quan Zou, Tianjin University, CHINA

**Received:** August 30, 2016

**Accepted:** November 14, 2016

**Published:** December 5, 2016

**Data Availability Statement:** All relevant data are within the paper and its Supporting Information files.

**Funding:** This work was partially supported by the National Natural Science Foundation of China (No. 11201409 to YL), the Young Talents Plan of Higher School in Hebei Province (No. BJ2014060 to YL) and the National Science Foundation of China (No 11171088 to YZ), the Science and technology project of Hebei Province (No A2015208108 and No 1520341 to YZ), the Science Fund of the Hebei University of Science and Technology Foundation (No 2014PT67 to YZ), the Hebei Province

## Abstract

In this paper, we have proposed a novel alignment-free method for comparing the similarity of protein sequences. We first encode a protein sequence into a 440 dimensional feature vector consisting of a 400 dimensional Pseudo-Markov transition probability vector among the 20 amino acids, a 20 dimensional content ratio vector, and a 20 dimensional position ratio vector of the amino acids in the sequence. By evaluating the Euclidean distances among the representing vectors, we compare the similarity of protein sequences. We then apply this method into the ND5 dataset consisting of the ND5 protein sequences of 9 species, and the F10 and G11 datasets representing two of the xylanases containing glycoside hydrolase families, i.e., families 10 and 11. As a result, our method achieves a correlation coefficient of 0.962 with the canonical protein sequence aligner ClustalW in the ND5 dataset, much higher than those of other 5 popular alignment-free methods. In addition, we successfully separate the xylanases sequences in the F10 family and the G11 family and illustrate that the F10 family is more heat stable than the G11 family, consistent with a few previous studies. Moreover, we prove mathematically an identity equation involving the Pseudo-Markov transition probability vector and the amino acids content ratio vector.

## Introduction

With the recent development of next-generation sequencing technologies, there has been an explosion in the numbers of available DNA and protein sequences. The numerous newly sequenced protein sequences present an urgent need for novel computational algorithms to compare their similarities with sequences from known protein families, to predict their structures, and thus to infer their functions [1–6].

As usually the first step in a bioinformatics pipeline, sequence comparison is very crucial since it affects all down-stream analyses. Popular methods for sequence comparison generally

Foundation for Advanced Talents (No A201400121 to YZ), the Educational Commission of Hebei Province on of Humanities and Social Sciences(No SZ16180 to YZ). 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.

fall into two categories: those using sequence alignment and those using alignment-free methods. In a sequence alignment, a score function is used to represent insertion, deletion, and substitution of nucleotides or amino acids in the compared DNAs or proteins, and the objective is to identity the alignment with the highest overall alignment score through methods like dynamic programming and seeding [7–9]. However, sometimes alignment becomes misleading due to unequal lengths of sequences, gene rearrangements, inversion, transposition, and translocation at substring level [10]. In these scenarios, alignment-free methods present good alternatives to alignment methods, which usually quantify sequence similarities using K-mer frequencies and other sequence features [11].

An alignment-free method for comparing protein sequences usually consists of two steps. At first, the protein sequences are transformed into fixed-length feature vectors [12]. The feature vectors are then fed into a vector similarity comparison algorithm to perform downstream analysis like phylogenetic inference [13]. Feature extraction is a procedure to extract desired information from the query sequences, which is usually critical to the accuracy of an alignment-free method [14]. Widely accepted features include chemical and physical properties [15], distance frequency matrix [16], K-string dictionary [17], 2D and 3D amino acid adjacency matrices [18], pseudo amino acid composition [19], and sequential and structural evolution information [20, 21]. Though these methods have their own advantages, they are suffering problems like computational intensive and low accuracy. Thus, more discriminatory features are still in demanding.

To further improve protein sequence comparison accuracy, we present a novel 440 dimensional feature vector, which models a few important information of a protein including the amino acids' abundance and position information, and the Pseudo-Markov transition probabilities among them. We then test the performance of our feature vector in two well studied datasets: (1) the ND5 dataset [22] and (2) the F10 and G11 dataset [23]. They have been widely used in evaluating protein comparison algorithms [22, 24]. As a result, our method is more accurate than a few existing methods for similarity analysis on the ND5 dataset, and we achieve accurate phylogenetic tree and heat stability results on the F10 and G11 dataset.

## Method

Amino acid composition and distribution are two most fundamental information about a protein sequence. They have been widely used and proven to be effective in protein sequence analyses [25], structural classification [26–28], pattern recognition receptor prediction [29], and fold recognition [30]. Thus, we proposed a novel representation for a protein sequence based on the two features, i.e. a 440-D feature vector consisting of (1) a 400-D Pseudo-Markov transition probability vector reflecting the order information of adjacent amino acids. (2) a 20-D amino acid content ratio vector describing the frequency of each amino acid in the sequence, and (3) a 20-D amino acid position ratio vector exhibiting the position distribution of each amino acid.

### Construction of the 400 dimensional Pseudo-Markov transition probability vector

Let $S = S_1S_2\cdots S_N$ be a protein sequence of length $N$ defined on $A = \{A_1, A_2, \cdots, A_{20}\}$, an ordered alphabet of 20 amino acids. For $1 \leq i, j \leq 20$, $1 \leq k \leq N$ and $1 \leq l \leq N - 1$, an amino acid $A_i$ is said to occur at position $k$ if $S_k = A_i$, and an ordered amino acid pair $A_iA_j$ is said to occur at position $l$ if $S_lS_{l+1} = A_iA_j$. Let $n_i$ be the number of occurrences of $A_i$ and $n_{i,j}$ be the number of occurrences of $A_iA_j$ in $S$. We then define the 400 dimensional vector as $(P_{1,1}, P_{1,2}, \cdots,$

$P_{1,20}, P_{2,1}, P_{2,2}, \cdots, P_{2,20}, \cdots, P_{20,1}, P_{20,2}, \cdots, P_{20,20}$), where

$$P_{i,j} = \begin{cases} \dfrac{n_{i,j}}{n_i} & \text{if } A_i \neq S_N \\ \dfrac{n_{i,j}}{n_i - 1} & \text{if } A_i = S_N \end{cases} \tag{1}$$

In particular, if there is no $A_i$ or $A_i$ appears only once at the end of $S$, then the numerator and denominator of $P_{i,j}$ are both 0. In this case, we define $P_{i,j} = 0$.

By definition, we have

$$\sum_{j=1}^{20} n_{i,j} = \begin{cases} n_i & \text{if } A_i \neq S_N \\ n_i - 1 & \text{if } A_i = S_N \end{cases} \tag{2}$$

$$\sum_{i=1}^{20} n_{i,j} = \begin{cases} n_j & \text{if } A_j \neq S_1 \\ n_j - 1 & \text{if } A_j = S_1 \end{cases} \tag{3}$$

From eqs (1) and (2) we have $\sum_{j=1}^{20} P_{i,j} = 1$, and thus $P_{i,j}$ can be considered as a transition probability from amino acid $A_i$ to $A_j$ in the protein sequence. So we call the 400 dimensional vector ($P_{1,1}, \ldots, P_{1,20}, P_{2,1}, \ldots, P_{2,20}, \ldots, P_{20,1}, \ldots, P_{20,20}$) a Pseudo-Markov transition probability vector.

## Construction of the 20 dimensional amino acid content ratio vector

Given that the protein sequence is composed of only 20 amino acids, it is clear that $\sum_{i=1}^{20} n_i = N$. For each amino acid $A_i$ ($1 \leq i \leq 20$), we define its content ratio $C_i$ as $C_i = \frac{n_i}{N}$ and the 20 dimensional amino acid content ratio vector as ($C_1, C_2, \ldots, C_{20}$). Obviously, $\sum_{i=1}^{20} C_i = 1$.

## Construction of the 20 dimensional amino acid position ratio vector

For each amino acid $A_i$ ($1 \leq i \leq 20$), let $s_i$ be the sum of all positions in $S$ that $A_i$ occurs. Noticing that $\sum_{i=1}^{20} s_i = \frac{N(N+1)}{2}$, we define the position ratio of the amino acid $D_i$ as $D_i = \frac{2s_i}{N(N+1)}$ and the 20 dimensional amino acid position ratio vector as ($D_1, D_2, \ldots, D_{20}$). Obviously, $\sum_{i=1}^{20} D_i = 1$.

By concatenating the above three types of vectors, we obtain a 440-D feature vector of $S$, that is, $V_s = (P_{1,1}, \ldots, P_{20,20}, C_1, \ldots, C_{20}, D_1, \ldots, D_{20})$. In the following, we show an interesting property of $V_s$. For $1 \leq j \leq 20$, let $\Delta_j = \sum_{i=1}^{20} C_i P_{i,j}$.

**Property.** Suppose $S_1 = A_u$ and $S_N = A_v$ for indices $u$ and $v$ with $1 \leq u, v \leq 20$. Then for any $1 \leq j \leq 20$, we have

$$\Delta_j = \begin{cases} C_j - \frac{1}{N} + \frac{P_{v,j}}{N} & \text{if } j = u \\ C_j + \frac{P_{v,j}}{N} & \text{if } j \neq u \end{cases}$$

In particular, if $A_v$ occurs only once in $S$, i.e. $n_v = 1$ then

$$\Delta_j = \begin{cases} C_j - \frac{1}{N} & \text{if } j = u \\ C_j & \text{if } j \neq u \end{cases}$$

**Proof.** If $j = u$, from eqs (1) and (3) we have

$$\begin{aligned} \Delta_u &= \sum_{i=1}^{20} C_i P_{i,u} = C_1 P_{1,u} + C_2 P_{2,u} + \cdots + C_u P_{u,u} + \cdots + C_v P_{v,u} + \cdots + C_{20} P_{20,u} \\ &= \frac{n_1}{N} \times \frac{n_{1,u}}{n_1} + \frac{n_2}{N} \times \frac{n_{2,u}}{n_2} + \cdots + \frac{n_u}{N} \times \frac{n_{u,u}}{n_u} + \cdots + \frac{n_v}{N} \times \frac{n_{v,u}}{n_v - 1} + \cdots + \frac{n_{20}}{N} \times \frac{n_{20,u}}{n_{20}} \\ &= \frac{n_{1,u}}{N} + \frac{n_{2,u}}{N} + \cdots + \frac{n_{u,u}}{N} + \cdots + \frac{n_v}{N} \times \frac{n_{v,u}}{n_v - 1} + \cdots + \frac{n_{20,u}}{N} \\ &= \frac{n_{1,u} + n_{2,u} + \cdots + n_{u,u} + \cdots + n_{v,u} + \cdots + n_{20,u} - n_{v,u}}{N} + \frac{n_v}{N} \times \frac{n_{v,u}}{n_v - 1} \\ &= \frac{n_u - 1 - n_{v,u}}{N} + \frac{n_v}{N} \times \frac{n_{v,u}}{n_v - 1} \\ &= \frac{n_u}{N} - \frac{1}{N} + \frac{n_{v,u}}{N} \left( \frac{n_v}{n_v - 1} - 1 \right) \\ &= C_u - \frac{1}{N} + \frac{P_{v,u}}{N} \end{aligned}$$

If $j \neq u$, we have

$$\begin{aligned} \Delta_j &= \sum_{i=1}^{20} C_i P_{i,j} = C_1 P_{1,j} + C_2 P_{2,j} + \cdots + C_u P_{u,j} + \cdots + C_v P_{v,j} + \cdots + C_{20} P_{20,j} \\ &= \frac{n_1}{N} \times \frac{n_{1,j}}{n_1} + \frac{n_2}{N} \times \frac{n_{2,j}}{n_2} + \cdots + \frac{n_u}{N} \times \frac{n_{u,j}}{n_u} + \cdots + \frac{n_v}{N} \times \frac{n_{v,j}}{n_v - 1} + \cdots + \frac{n_{20}}{N} \times \frac{n_{20,j}}{n_{20}} \\ &= \frac{n_{1,j}}{N} + \frac{n_{2,j}}{N} + \cdots + \frac{n_{u,j}}{N} + \cdots + \frac{n_v}{N} \times \frac{n_{v,j}}{n_v - 1} + \cdots + \frac{n_{20,j}}{N} \\ &= \frac{n_{1,j} + n_{2,j} + \cdots + n_{u,j} + \cdots + n_{v,j} + \cdots + n_{20,j} - n_{v,j}}{N} + \frac{n_v}{N} \times \frac{n_{v,j}}{n_v - 1} \\ &= \frac{n_j - n_{v,j}}{N} + \frac{n_v}{N} \times \frac{n_{v,j}}{n_v - 1} \\ &= \frac{n_j}{N} + \frac{n_{v,j}}{N} \left( \frac{n_v}{n_v - 1} - 1 \right) \\ &= C_j + \frac{P_{v,j}}{N} \end{aligned}$$

Finally, let $n_v = 1$. By definition, we have $n_{v,j} = 0$ and $P_{v,j} = 0$ for any $1\leq j \leq 20$. Thus,

$$\Delta_j = \begin{cases} C_j - \dfrac{1}{N} & \text{if } j = u \\ C_j & \text{if } j \neq u \end{cases},$$

completing the proof.

## Quantifying the distances among protein sequences based on their feature vectors

Let $S$ and $T$ be two proteins and $V_S$ and $V_T$ be their 440-D feature vectors. Then the distance between $S$ and $T$ is quantified by the Euclidean distance between $V_S$ and $V_T$, that is, $d(S,T) = \sqrt{\sum_{i=1}^{440} (V_S[i] - V_T[i])^2}$, where $V_S[i]$ and $V_T[i]$ denote the i$^{\text{th}}$ entries of the vectors $V_S$ and $V_T$ respectively.

## Results and Discussions

To evaluate the performance of our method, we applied it into two datasets: (1) the ND5 dataset [22] and (2) the F10 and G11 dataset [23].

## Datasets

The ND5 dataset consists of the ND5 protein sequences of 9 species including human, gorilla, pigmy chimpanzee, common chimpanzee, fin whale, blue whale, rat, mouse, and opossum (Table 1). The sequences have lengths 602~610 base pairs (bps). It is a popular benchmark data for testing the performances of computational methods in comparing the similarity of protein sequences [15, 31–34].

The F10 and G11 datasets represent two of the xylanases containing glycoside hydrolase families, i.e., families 10 and 11 respectively. Specifically, the F10 dataset contains ten xylanases with NCBI accession IDs O59859, P56588, P33559, Q00177, P07986, P07528, P40943, P23556, P45703, and Q60041 respectively. The G11 dataset also consists of ten xylanases with NCBI IDs P33557, P55328, P55331, P45705, P26220, P55334, Q06562, P55332, P55333, and P17137 respectively.

**Table 1. Information of ND5 for nine species.**

| Species | ID (NCBI) | Length |
|---|---|---|
| Human (Homo sapiens) | AP_000649 | 603 |
| Gorilla (Gorilla gorilla) | NP_008222 | 603 |
| Pigmy chimpanzee (Pan paniscus) | NP_008209 | 603 |
| Common chimpanzee (Pan troglodytes) | NP_008196 | 603 |
| Fin whale (Balenoptera physalus) | NP_006899 | 606 |
| Blue whale (Balenoptera musculus) | NP_007066 | 606 |
| Rat (Tattus norvegicus) | NP_004902 | 610 |
| Mouse (Mus musculus) | NP_904338 | 607 |
| Opossum (Didelphis virginiana) | NP_007105 | 602 |

doi:10.1371/journal.pone.0167430.t001

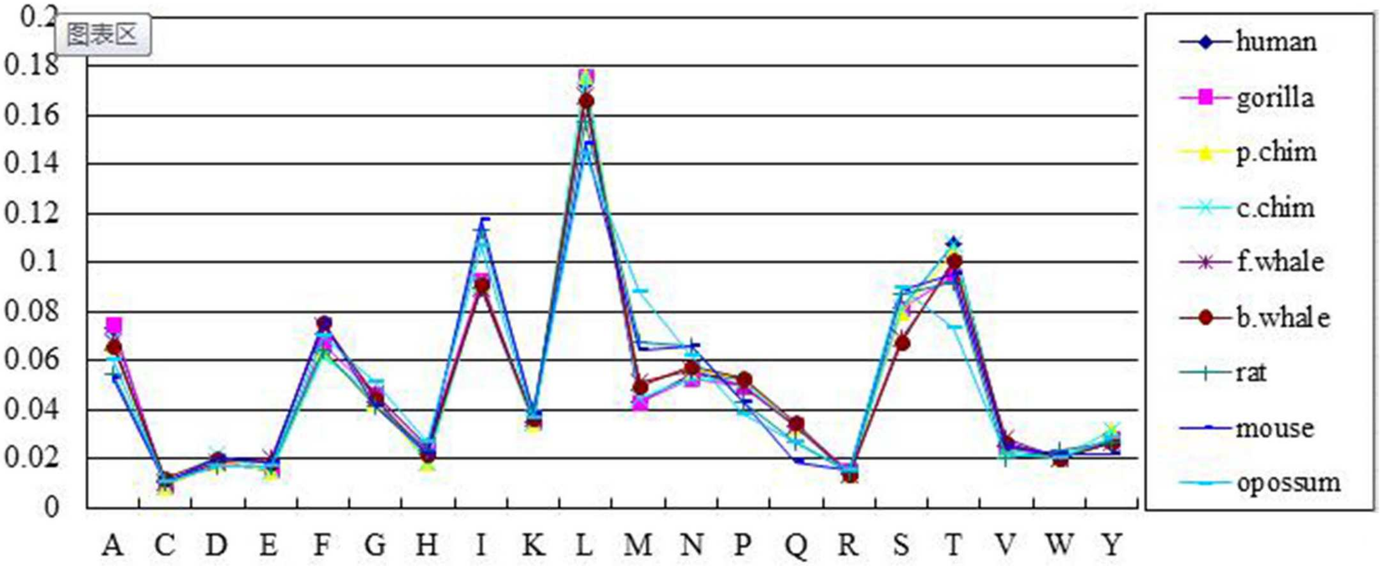

**Fig 1. The content ratios of twenty amino acids in the ND5 dataset.** The X axis denotes the 20 amino acids and the Y axis denotes the content ratios of each amino acid for the 9 sequences.

doi:10.1371/journal.pone.0167430.g001

## Application to the ND5 dataset

We first encoded the nine protein sequences into 440-D feature vectors. In Figs 1 and 2, we showed the content ratios and position ratios of the twenty amino acids over the sequences.

As can be seen, the content ratios and position ratios exhibit similar trends over the twenty amino acids. The amino acid L has the highest content ratio and position ratio over all 9 sequences whereas amino acid C has the lowest content ratio and position ratio. In addition, the 9 species are quite similar according to the amino acids distributions of both the content ratio and position ratio in the ND5 protein.

We then calculated the pairwise Euclidean distances among the nine 440-D feature vectors and showed the results in Table 2. As we can see, human, P.chim, C.chim, and gorilla are closer to each other and they are relatively far from rat, mouse and opossum. For a better view, we also plotted a heat-map based on the distances in Fig 3.

In order to estimate the contribution of each part in the 440-D feature vector to the final performance in sequence similarity analysis, we plotted heat maps for the ND5 dataset based on the 20-D amino acid position ratio vector (see Fig 4), the 20-D amino acid content ratio vector (see Fig 5), and the 40-D amino acid position ratio and content ratio vector (see Fig 6), respectively. Clearly, Fig 3 is most consistent with the known result from the 440-D vector, Fig 6 is a little bit worse, and Figs 4 and 5 are the worst. As an indication, the 400-D Pseudo-Markov transition probability vector plays major role in sequence comparison.

A common strategy to evaluate an alignment-free method is to compare it with a popular alignment method like ClustalW [31], which has a much higher time and space complexity than alignment-free methods. Table 3 showed the pair-wise distances of the 9 protein sequences using ClustalW (i.e. Table 4 in [31]). We calculated the correlation coefficient between the distances from our method and those from ClustalW and compared our method with a few popular alignment-free methods [15, 31–34] using this coefficient as a criterion (see Table 4).

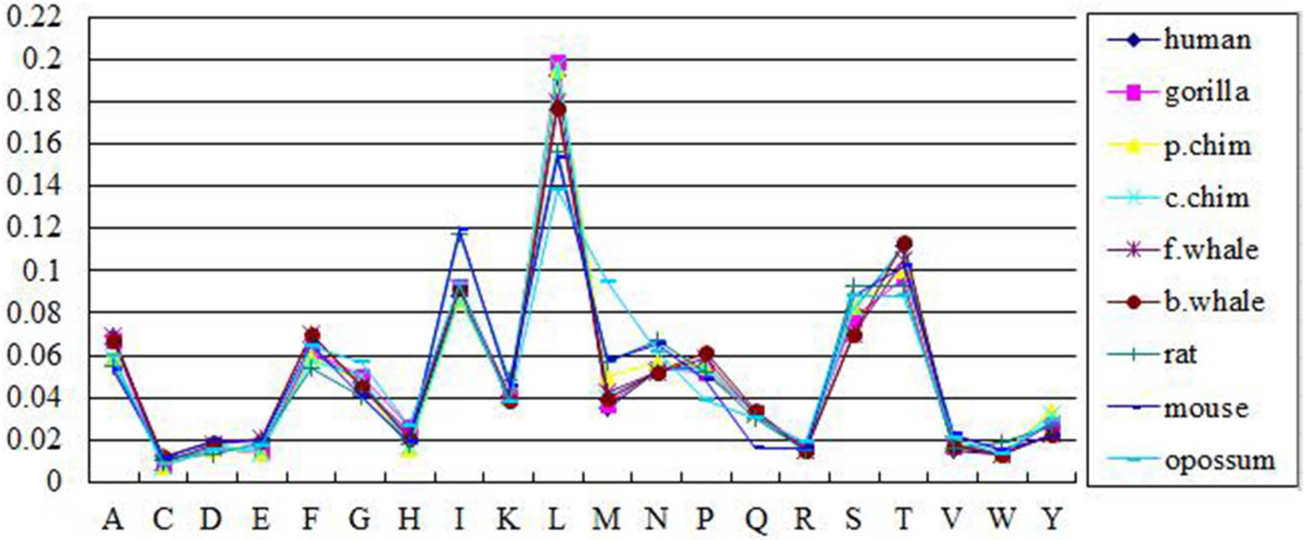

**Fig 2. The position ratios of twenty amino acids in the ND5 dataset.** The X axis denotes the 20 amino acids and the Y axis denotes the position ratios of each amino acid for 9 sequences.

doi:10.1371/journal.pone.0167430.g002

As Table 4 shows, the correlation coefficient between our method and ClustalW is 0.962, which is the highest among the 6 methods. As a result, our method is more consistent with ClustalW than the other 5 methods, which indicates that our method is more accurate.

## Application to the F10 and G11 dataset

We also tested our method on the F10 and G11 datasets and plotted the heat map based on the pair-wise Euclidean distances in Fig 7. As can be seen, our method accurately separated the sequences in family F10 with those in G11 with the F10 xylanases locating in the top right quarter and G11 xylanases in the lower left quarter. We also observed that the F10 dataset is more heat stable than the G11 dataset, which is consistent with other studies, e.g.,[15].

It is of note that we applied the Euclidean distance in quantifying the distances among the feature vectors for different proteins. Euclidean distance is one of the simplest and most intuitive distance measures, which has been adopted in many fields, such as gene identification [35], protein 3D structure reconstruction [36], robust automatic pectoral muscle segmentation [37] and classification of normal and epileptic seizure EEG signals [38], etc. However, there are many other distance measures, which could affect protein similarity analysis. As an example, we compared the Euclidean distance with the Hamming distance for the ND5 dataset and

**Table 2. The distance matrix of nine species by our method.**

| | Human | Gorilla | P.chim | C.chim | F.whale | B.whale | Rat | Mouse | Opossum |
|---|---|---|---|---|---|---|---|---|---|
| **Human** | 0 | 0.53 | 0.497 | 0.501 | 0.764 | 0.782 | 0.901 | 0.945 | 0.972 |
| **Gorilla** | | 0 | 0.522 | 0.564 | 0.755 | 0.803 | 0.876 | 0.963 | 1.025 |
| **P.chim** | | | 0 | 0.381 | 0.743 | 0.742 | 0.88 | 0.913 | 0.922 |
| **C.chim** | | | | 0 | 0.766 | 0.773 | 0.903 | 0.949 | 0.981 |
| **F.whale** | | | | | 0 | 0.347 | 0.868 | 0.906 | 1.012 |
| **B.whale** | | | | | | 0 | 0.914 | 0.898 | 0.990 |
| **Rat** | | | | | | | 0 | 0.74 | 0.955 |
| **Mouse** | | | | | | | | 0 | 0.859 |

doi:10.1371/journal.pone.0167430.t002

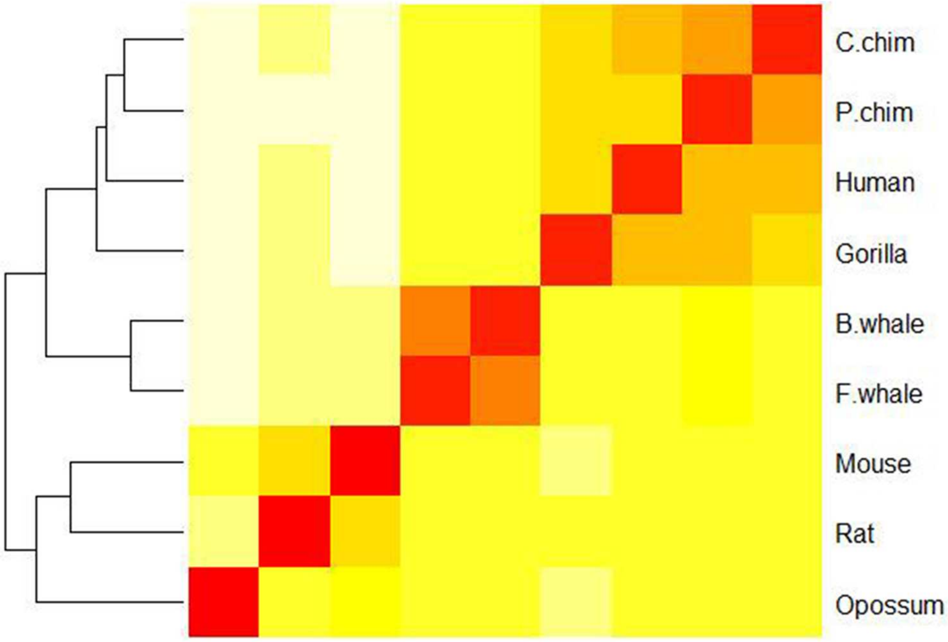

**Fig 3. A heat map showing the similarity of nine species in the ND5 dataset.** Red color indicates small distance and high similarity between the sequences and yellow color indicates large distance and low similarity, the same as below.

doi:10.1371/journal.pone.0167430.g003

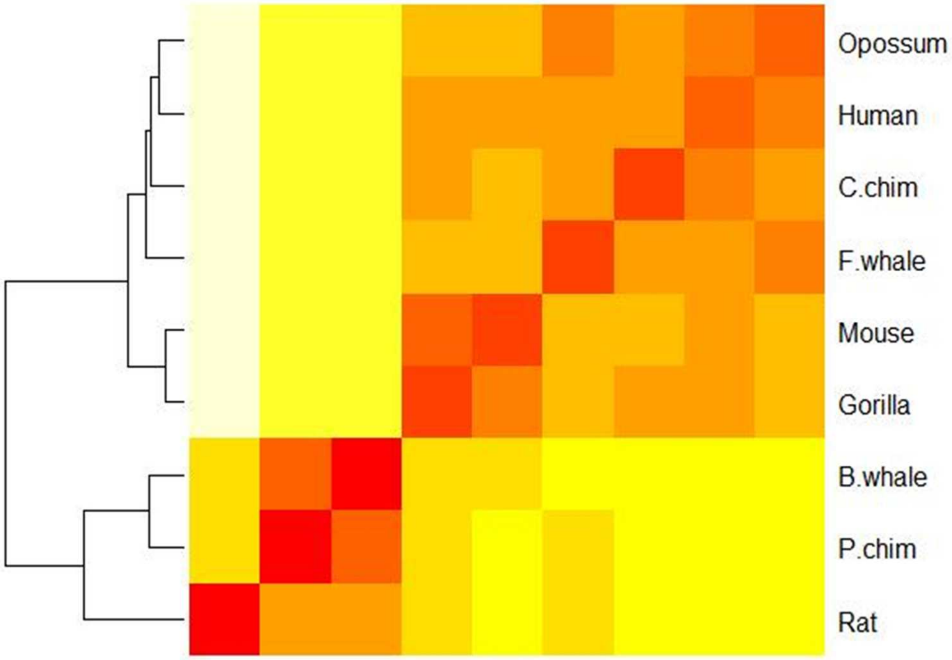

**Fig 4. A heat map showing the similarity of nine species in the ND5 dataset based on the 20-D amino acid position ratio vector.**

doi:10.1371/journal.pone.0167430.g004

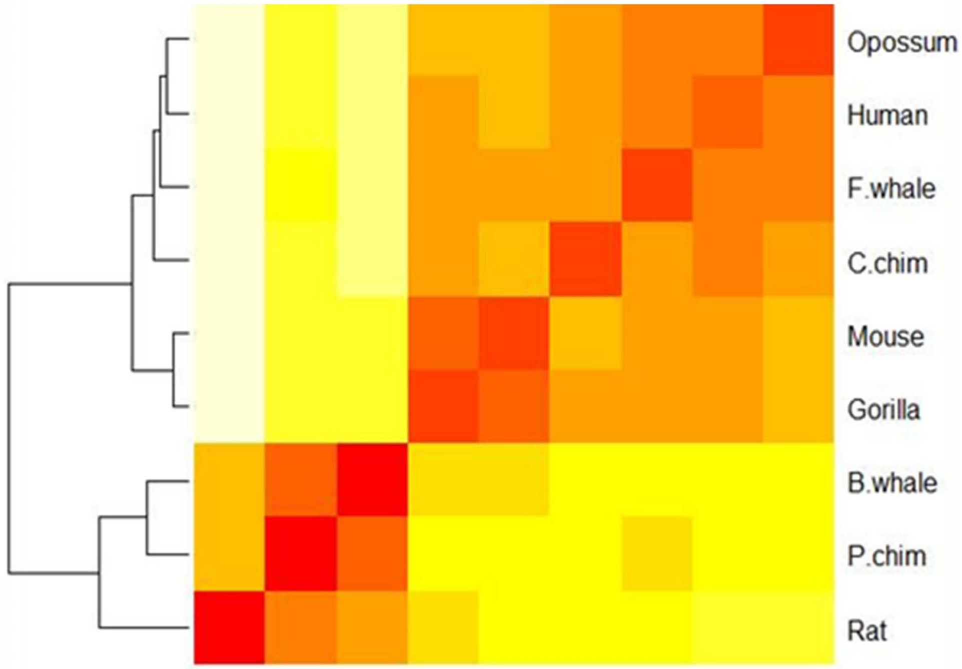

**Fig 5. A heat map showing the similarity of nine species in the ND5 dataset based on the 20-D amino acid content ratio vector.**

doi:10.1371/journal.pone.0167430.g005

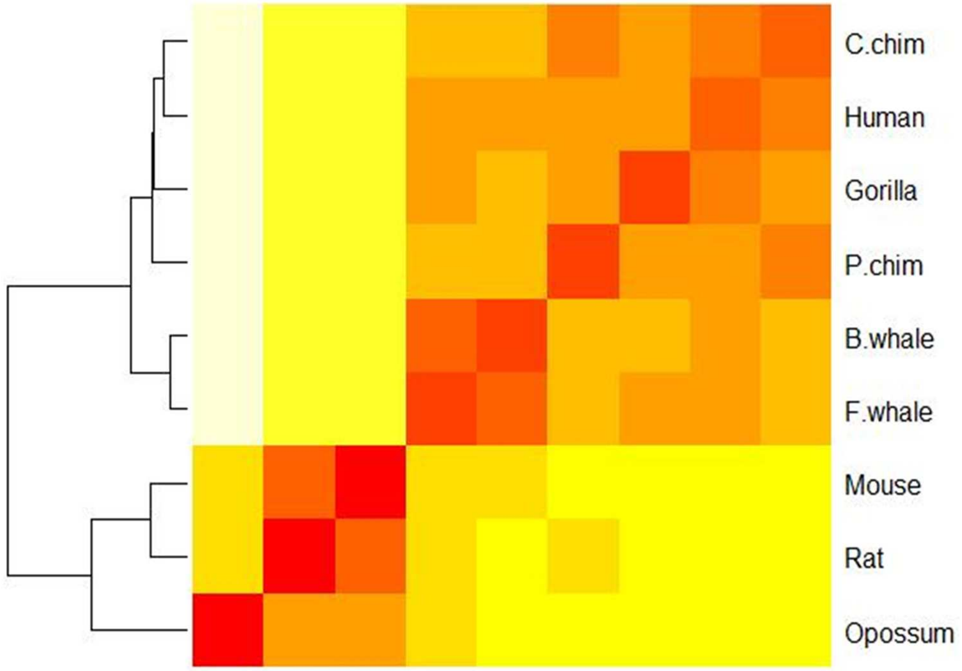

**Fig 6. A heat map showing the similarity of nine species in the ND5 dataset based on the 40-D amino acid position ratio and content ratio vector.**

doi:10.1371/journal.pone.0167430.g006

Table 3. The distance matrix of nine species calculated by ClustalW (i.e. Table 4 in [31]).

| | Human | Gorilla | P.chim | C.chim | F.whale | B.whale | Rat | Mouse | Opossum |
|---|---|---|---|---|---|---|---|---|---|
| **Human** | 0 | 10.7 | 7.1 | 6.9 | 41.0 | 41.3 | 50.2 | 48.9 | 50.4 |
| **Gorilla** | | 0 | 9.7 | 9.9 | 42.7 | 42.4 | 51.4 | 49.9 | 54.0 |
| **P.chim** | | | 0 | 5.1 | 40.1 | 40.1 | 50.2 | 48.9 | 50.1 |
| **C.chim** | | | | 0 | 40.4 | 40.4 | 50.8 | 49.6 | 51.4 |
| **F.whale** | | | | | 0 | 3.5 | 45.3 | 46.8 | 52.7 |
| **B.whale** | | | | | | 0 | 45.0 | 45.9 | 52.7 |
| **Rat** | | | | | | | 0 | 25.9 | 54.0 |
| **Mouse** | | | | | | | | 0 | 50.8 |

doi:10.1371/journal.pone.0167430.t003

Table 4. Comparison of 6 alignment-free methods.

| Method | Correlation coefficients |
|---|---|
| Our method | 0.962 |
| Ma et al. [31] (Table 3*) | 0.9304 |
| Matty et al. [32] (Table 3*) | 0.6594 |
| Bielińska-Wąż [34] (Table 4*) | 0.7280 |
| Wen et al. [33] (Table 3*) | 0.7324 |
| Yao et al. [15] (Table 3*) | 0.6908 |

*The table in the literatures listed the correlation coefficient between the corresponding method and ClustalW.

doi:10.1371/journal.pone.0167430.t004

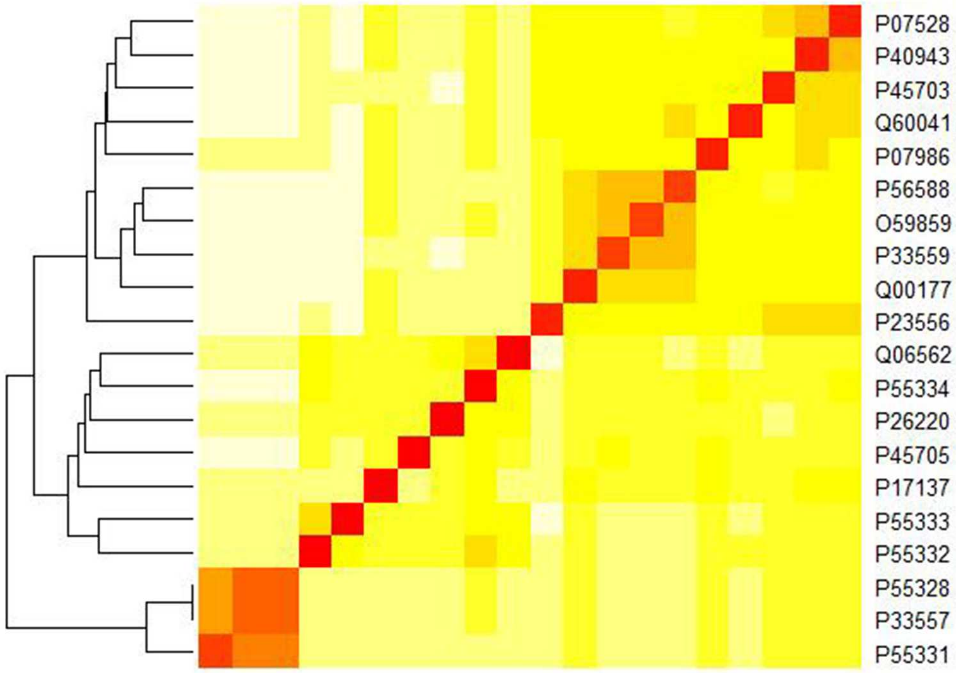

Fig 7. A heat map showing the similarity of 20 xylanases in the F10 and G11 datasets.

doi:10.1371/journal.pone.0167430.g007

F10 and G11 datasets respectively. We also plotted the heat-map for the ND5 dataset in S1 Fig based on the Hamming distance. Similar plots for the F10 and G11 datasets were shown in S2 Fig. Interestingly, Fig 3 and S1 Fig are almost the same while Fig 7 and S2 Fig exhibit significant differences. Clearly, Fig 7 (based on the Euclidean distance) is better since the two xylanases families are well separated while S2 Fig (based on the Hamming distance) fails to do it. For the ND5 dataset, we further computed the agreement (i.e., the Pearson correlation coefficients between the protein similarity matrices) between our method (based on the Hamming distance) and ClustalW, which is 0.937, a little bit less than that for the Euclidean distance (0.962). Thus, we believe that the Euclidean distance is more effective than Hamming distance for these two datasets.

## Conclusion

In this paper, we have proposed a novel alignment-free method to compare protein sequences. The method is more accurate than 5 other popular alignment-free methods in the ND5 dataset and is capable of distinguishing the F10 xylanases family from the G11 family. The comparison results of this method are quite consistent with protein sequence aligners like ClustalW. It presents an alternative of these aligners when time and space complexities become an issue.

In the future, a few machine learning methods [39] could be applied to further improve the performance of our method. For example, in contrast to phylogenetic analysis, methods like K-means analysis [40] and random forest [41] could also be applied to classify the proteins and perform taxonomy. However, it is out of the scope of this study. In addition, our novel features could also be applied into applications like essential gene identification [42] and similar problems related to DNAs or RNAs.

## Supporting Information

**S1 Fig. A heat map showing the similarity of nine species in the ND5 dataset based on the Hamming distance.**
(TIF)

**S2 Fig. A heat map showing the similarity of 20 xylanases in the F10 and G11 datasets based on the Hamming distance.**
(TIF)

**S1 Table. The nine ND5 protein sequences.**
(TXT)

**S2 Table. The 10 sequences in the F10 xylanase family.**
(TXT)

**S3 Table. The 10 sequences in the G11 xylanase family.**
(TXT)

## Author Contributions

**Conceptualization:** YL.

**Data curation:** YZ JLY.

**Formal analysis:** TS.

**Funding acquisition:** YL YZ.

**Investigation:** JSY YZ.

**Methodology:** YL JLY.

**Project administration:** YL.

**Resources:** YL.

**Software:** TS.

**Supervision:** JSY YZ.

**Validation:** YL JLY.

**Visualization:** YL JLY.

**Writing – original draft:** JLY YL TS.

**Writing – review & editing:** YL YZ.

## References

1. Zhang L, Zhao X, Kong L. Predict protein structural class for low-similarity sequences by evolutionary difference information into the general form of Chou[U+05F3]s pseudo amino acid composition. Journal of Theoretical Biology. 2014;355.
2. Zhang S, Liang Y, Yuan X. Improving the prediction accuracy of protein structural class: Approached with alternating word frequency and normalized Lempel–Ziv complexity. Journal of Theoretical Biology. 2014; 341(1):71–7.
3. Wang J, Yan L, Liu X, Qi D, Yao Y, He P. High-accuracy Prediction of Protein Structural Classes Using PseAA Structural Properties and Secondary Structural Patterns. Biochimie. 2014; 101(6):104–12.
4. Liang K, Zhang L, Lv J. Accurate prediction of protein structural classes by incorporating predicted secondary structure information into the general form of Chou's pseudo amino acid composition. Journal of Theoretical Biology. 2014; 344:12–8. doi: 10.1016/j.jtbi.2013.11.021 PMID: 24316044
5. Xiao X, Shao SH, Huang ZD, Chou KC. Using pseudo amino acid composition to predict protein structural classes: approached with complexity measure factor. Journal of Computational Chemistry. 2006; 27(4):478–82. doi: 10.1002/jcc.20354 PMID: 16429410
6. Gu Q, Ding YS, Zhang TL. Prediction of G-Protein-Coupled Receptor Classes in Low Homology Using Chou's Pseudo Amino Acid Composition with Approximate Entropy and Hydrophobicity Patterns. Protein & Peptide Letters. 2010; 17(5):559–67.
7. Smith TF, Waterman MS. Identification of common molecular subsequences. Journal of molecular biology. 1981; 147(1):195–7. PMID: 7265238
8. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. Journal of molecular biology. 1990; 215(3):403–10. doi: 10.1016/S0022-2836(05)80360-2 PMID: 2231712
9. Yang J, Zhang L. Run probabilities of seed-like patterns and identifying good transition seeds. Journal of computational biology: a journal of computational molecular cell biology. 2008; 15(10):1295–313.
10. Otu HH, Sayood K. A new sequence distance measure for phylogenetic tree construction. Bioinformatics. 2003; 19(16):2122–30. PMID: 14594718
11. Zhang Y, Huang H, Dong X, Fang Y, Wang K, Zhu L, et al. A Dynamic 3D Graphical Representation for RNA Structure Analysis and Its Application in Non-Coding RNA Classification. PloS one. 2016; 11(5): e0152238. doi: 10.1371/journal.pone.0152238 PMID: 27213271
12. Yao Y, Yan S, Han J, Dai Q, He PA. A novel descriptor of protein sequences and its application. Journal of Theoretical Biology. 2014; 347(4):109–17.
13. Liao B, Shan X, Zhu W, Li R. Phylogenetic tree construction based on 2D graphical representation. Chemical Physics Letters. 2006; 422(s 1–3):282–8.
14. Nandy A, Harle M, Basak SC. Mathematical descriptors of DNA sequences: Development and application. Arkivoc. 2006; 2006(IX):211–38.
15. Yao Y, Dai Q, C, He P, Nan X, Zhang Y. Analysis of similarity/dissimilarity of protein sequences. Proteins Structure Function & Bioinformatics. 2008; 73(4):864–71.
16. Mu Z, Wu J, Zhang Y. A novel method for similarity/dissimilarity analysis of protein sequences. Physica A Statistical Mechanics & Its Applications. 2013; 392(24):6361–6.

17. Y Chenglong, He RL, Y Stephen S-T. Protein sequence comparison based on K-string dictionary. Gene. 2013; 529(2):250–6. doi: 10.1016/j.gene.2013.07.092 PMID: 23939466

18. El-Lakkani A, El-Sherif S. Similarity analysis of protein sequences based on 2D and 3D amino acid adjacency matrices. Chemical Physics Letters. 2013; 590(12):192–5.

19. Yu HJ, Huang DS. Novel 20-D descriptors of protein sequences and it's applications in similarity analysis. Chemical Physics Letters. 2012; 531:261–6.

20. Wei L, Liao M, Gao X, Zou Q. Enhanced Protein Fold Prediction Method Through a Novel Feature Extraction Technique. IEEE Transactions on Nanobioscience. 2015; 14(6):649–59. doi: 10.1109/TNB.2015.2450233 PMID: 26335556

21. Wei L, Liao M, Gao X, Zou Q. An Improved Protein Structural Prediction Method by Incorporating Both Sequence and Structure Information. Nanobioscience IEEE Transactions on. 2014; 34(4):545–59.

22. Liao B, Liao B, Sun X, Zeng Q. A novel method for similarity analysis and protein sub-cellular localization prediction. Bioinformatics (Oxford, England). 2010; 26(21):2678–83.

23. Collins T, Gerday C, Feller G. Xylanases, xylanase families and extremophilic xylanases. FEMS Microbiol Rev. 2005; 29(1):3–23. doi: 10.1016/j.femsre.2004.06.005 PMID: 15652973

24. Randic M, Mehulic K, Vukicevic D, Pisanski T, Vikic-Topic D, Plavsic D. Graphical representation of proteins as four-color maps and their numerical characterization. J Mol Graph Model. 2009; 27(5):637–41. doi: 10.1016/j.jmgm.2008.10.004 PMID: 19081277

25. Xu C, Sun D, Liu S, Zhang Y. Protein sequence analysis by incorporating modified chaos game and physicochemical properties into Chou's general pseudo amino acid composition. J Theor Biol. 2016; 406:105–15. doi: 10.1016/j.jtbi.2016.06.034 PMID: 27375218

26. Zhang L, Zhao X, Kong L. Predict protein structural class for low-similarity sequences by evolutionary difference information into the general form of Chou's pseudo amino acid composition. J Theor Biol. 2014; 355:105–10. doi: 10.1016/j.jtbi.2014.04.008 PMID: 24735902

27. Zhang S, Ye F, Yuan X. Using principal component analysis and support vector machine to predict protein structural class for low-similarity sequences via PSSM. Journal of biomolecular structure & dynamics. 2012; 29(6):634–42.

28. Kong L, Zhang L, Lv J. Accurate prediction of protein structural classes by incorporating predicted secondary structure information into the general form of Chou's pseudo amino acid composition. J Theor Biol. 2014; 344:12–8. doi: 10.1016/j.jtbi.2013.11.021 PMID: 24316044

29. Gao QB, Zhao H, Ye X, He J. Prediction of pattern recognition receptor family using pseudo-amino acid composition. Biochemical and biophysical research communications. 2012; 417(1):73–7. doi: 10.1016/j.bbrc.2011.11.057 PMID: 22138239

30. Ding CH, Dubchak I. Multi-class protein fold recognition using support vector machines and neural networks. Bioinformatics. 2001; 17(4):349–58. PMID: 11301304

31. Ma T, Liu Y, Dai Q, Yao Y, He PA. A graphical representation of protein based on a novel iterated function system. Physica A Statistical Mechanics & Its Applications. 2014; 403(6):21–8.

32. Maaty MIAE, Abo-Elkhier MM, Elwahaab MAA. 3D graphical representation of protein sequences and their statistical characterization. Physica A Statistical Mechanics & Its Applications. 2010; 389 (21):4668–76.

33. Wen J, Zhang YY. A 2D graphical representation of protein sequence and its numerical characterization. Chemical Physics Letters. 2009; 476(4):281–6.

34. Bielińska-Wąż D. Graphical and numerical representations of DNA sequences: statistical aspects of similarity. Journal of Mathematical Chemistry. 2011; 49(10):2345–407.

35. Ghosh A, Barman S. Application of Euclidean distance measurement and principal component analysis for gene identification. Gene. 2016; 583(2):112–20. doi: 10.1016/j.gene.2016.02.015 PMID: 26877227

36. Pietal MJ, Bujnicki JM, Kozlowski LP. GDFuzz3D: a method for protein 3D structure reconstruction from contact maps, based on a non-Euclidean distance function. Bioinformatics. 2015; 31(21):3499–505. doi: 10.1093/bioinformatics/btv390 PMID: 26130575

37. Bora VB, Kothari AG, Keskar AG. Robust Automatic Pectoral Muscle Segmentation from Mammograms Using Texture Gradient and Euclidean Distance Regression. J Digit Imaging. 2016; 29(1):115–25. doi: 10.1007/s10278-015-9813-5 PMID: 26259521

38. Lee SH, Lim JS, Kim JK, Yang J, Lee Y. Classification of normal and epileptic seizure EEG signals using wavelet transform, phase-space reconstruction, and Euclidean distance. Comput Methods Programs Biomed. 2014; 116(1):10–25. doi: 10.1016/j.cmpb.2014.04.012 PMID: 24837641

39. Wei L, Liao M, Gao Y, Ji R, He Z, Zou Q. Improved and Promising Identification of Human MicroRNAs by Incorporating a High-Quality Negative Set. IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM. 2014; 11(1):192–201.

**40.** Dubey AK, Gupta U, Jain S. Analysis of k-means clustering approach on the breast cancer Wisconsin dataset. Int J Comput Assist Radiol Surg. 2016; 11(11):2033–47. doi: 10.1007/s11548-016-1437-9 PMID: 27311823

**41.** Liao Z, Ju Y, Zou Q. Prediction of G Protein-Coupled Receptors with SVM-Prot Features and Random Forest. Scientifica (Cairo). 2016; 2016:8309253.

**42.** Hua HL, Zhang FZ, Labena AA, Dong C, Jin YT, Guo FB. An Approach for Predicting Essential Genes Using Multiple Homology Mapping and Machine Learning Algorithms. BioMed research international. 2016; 2016:7639397. doi: 10.1155/2016/7639397 PMID: 27660763