BIOINFORMATICS ORIGINAL PAPER

Vol. 26 no. 21 2010, pages 2678–2683
doi:10.1093/bioinformatics/btq521

Sequence analysis Advance Access publication September 8, 2010

A Novel method for similarity analysis and protein sub-cellular localization prediction

Bo Liao*, Benyou Liao, Xingming Sun and Qingguang Zeng

School of computer and communication, Hunan University, Changsha Hunan, 410082, China

Associate Editor: John Quackenbush

ABSTRACT

Motivation: Biological sequence was regarded as an important study by many biologists, because the sequence contains a large number of biological information, what is helpful for scientists' studies on biological cells, DNA and proteins. Currently, many researchers used the method based on protein sequences in function classification, sub-cellular location, structure and functional site prediction, including some machine-learning methods. The purpose of this article, is to find a new way of sequence analysis, but more simple and effective.

Results: According to the nature of 64 genetic codes, we propose a simple and intuitive 2D graphical expression of protein sequences. And based on this expression we give a new Euclidean-distance method to compute the distance of different sequences for the analysis of sequence similarity. This approach contains more sequence information. A typical phylogenetic tree constructed based on this method proved the effectiveness of our approach. Finally, we use this sequence-similarity-analysis method to predict protein sub-cellular localization, in the two datasets commonly used. The results show that the method is reasonable.

Contact: dragonbw@163.com

Received on August 4, 2010; revised on August 26, 2010; accepted on September 3, 2010

1 INTRODUCTION

Bioinformatics is one of the great frontiers of life sciences, and it is also be one of the core areas of Natural Science in 21st century. Now, many researchers pay their attentions on protein function classification and sub-cellular localization. The methods of such studies, classified from the data characteristics, include sequence-based methods, based on interaction networks, functional domains and so on. The sequence-based method has two ways, homologous and non-homologous. The non-homologous method is constructing information from the sequence based on the chemical–physical characters, such as the amino acid composition, the frequencies of different sub-sequences, the interval numbers of one sub-sequences (Al-Shahib *et al*., 2005a, b; Gao *et al*., 2005; Lee *et al*., 2009; Li *et al*., 2009; Zhang *et al*., 2009) and so on. The homologous-based method is mainly analyzing the amino acid sequence similarity, including the way using the alignment and the way computing the distance of sequences, using the mathematical descriptors abstracted from some matrixes, based on some graphical representations. The graphical representation what do not need machine learning or other complicated computing, can provide intuitive picture or useful insights for helping analyzing.

Since 1983, many graphical representations of sequences have been provided in different biological topics, such as 2D-, 3D-, 4D-, 6D-graphical representations of DNA sequences (Cao *et al*., 2008; Liao and Zhu, 2006; Liao *et al*., 2005, 2007; Liu *et al*., 2009; Nandy, 1996; Randić *et al*., 2000a, b, 2003a, b; Yu and Sun, 2010; Yu *et al*., 2009). In recent years, many 2D graphical representations of proteins have been proposed by Randić and others (2004, 2009) (Bai and Wang, 2005; He *et al*., 2010; Randić, 2007; Wen and Zhang, 2009). Also, many condensed matrices were provided (Cao *et al*., 2008; Liao and Zhu, 2006; Liao *et al*., 2005, 2007; Liu *et al*., 2009; Nandy, 1996; Randić *et al*., 2000a, b, 2003a, b; Yu and Sun, 2010; Yu *et al*., 2003, 2009), such as, D/D matrix in which entries represent the quotient of the Euclidean and the graph-theoretical distance between vertices in 2D plane; L/L matrix whose elements are defined as the quotient of the Euclidean distance between a pair of vertices (dots) of curve and the sum of distances between the same pair of vertices; M, M/M, CM and so on. Based on these matrices, many invariants can be obtained for comparison of sequences. Applying the above methods, the researchers have compared the similarities and dissimilarities of sequences (Liao and Zhu, 2006). Also many researchers had used these matrices in constructing phylogenetic tree.

In this work, we propose a new Euclidean distance computing method based on a new graphical representation composed basing on genetic codes distribution. Then we apply this method in protein sub-cellular localization prediction using the similarity comparisons.

2 NEW GRAPHICAL REPRESENTATION OF PROTEIN SEQUENCE

As we all know that the four nucleic bases A, G, T and C can build up 64 kinds of trinucleotides. Based on the reason that the second base of a trinucleotide is associated with the hydrophobic/hydrophilic property of the translated amino acid, Jiafeng Yu (Yu *et al*., 2009) classified the 64 kinds of trinucleotides into four categories and proposed a 3D graphical representation based on trinucleotides. We translated the codes into amino acids presented in Figure 1.

From Figure 1, we can get that N, K, D, H, Y, E, Q are in the first quadrant; R, C, G, S, W are in the second quadrant; P, A, T, S are in the third quadrant; I, M, V, L, F are in the forth quadrant. Based on the principles of symmetry and the nearest, we can set the 20 amino acids in the four quadrants and the two axis of Cartesian 2D coordinate. We set three amino acids in every quadrant, and every axis with four amino acids, two in the positive side, while two in the negative side. S is in the second and third quadrants, so we set

*To whom correspondence should be addressed.

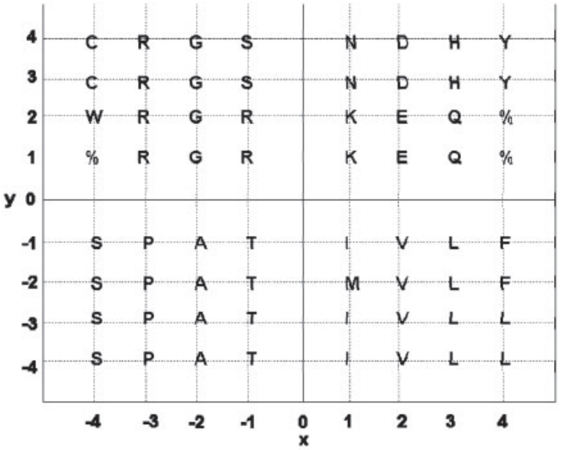

Fig. 1. Distributions of the amino acids in the four quadrants of Cartesian 2D coordinate.

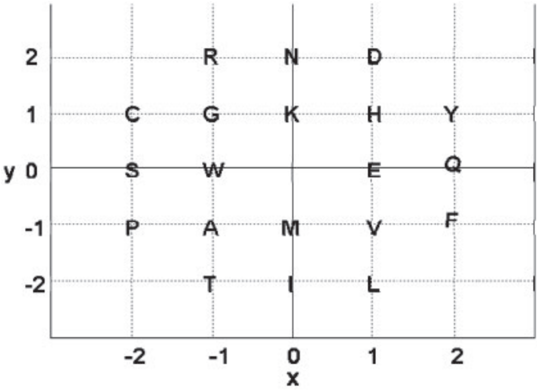

Fig. 2. The redistributions of the amino acids in the four quadrants of Cartesian 2D coordinate with no repeat.

it on the x-axis; N, K are near the y-axis in the first quadrant, so, it can be set on the positive axis of Y; M, I are also near the y-axis in the forth quadrant, so they take their places on the negative axis of y; E, Q are near the x-axis in the first quadrant, so they are set on the positive axis of x; W appears only once in the second quadrant, for the symmetry, we put it on the negative axis of x; The rest of the trinucleotide are set in the quadrants where they are in Figure 1. By this method of distribution of the amino acids, based on their distances and neighbor relations each trinucleotide can be represented by a set of coordinate (x, y). The 64 trinucleotide can be assigned as shown in Figure 2:

In detail, N = (0, 2), K = (0, 1), H = (1, 1), D = (1, 2), Y = (2, 1), R = (−1, 2), G = (−1, 1), W = (−1, 0), C = (−2, 1), S = (−2, 0), P = (−2, −1), A = (−1, −1), T = (−1, −2), M = (0, −1), I = (0, −2), L = (1, −2), V = (1, −1), E = (1, 0), Q = (2, 0), F = (2, −1).

For every trinucleotide, we can get another component z, $z = x \times y$; For a protein sequence $S = s_1\ s_2\ s_3\ s_4 \ldots s_N$, where N is the length of protein sequence, every amino acid can get a point as $s_i = (x_i, y_i, z_i)$. Letting $X_i = \sum_{n=1}^{i} x_n$; $Y_i = \sum i_{n=1}\ y_n$, $Z_i = \sum_{n=1}^{i} z_n$, so, every amino acid can be represented by an another point as $s'_i = (X_i, Y_i, Z_i)$.

Table 1. The 30 points value and the distance between the corresponding coordinate of the two protein sequences

Human	x	y	z	Common chimpanzee	x	y	z	Distance	Cumulative
M	0	−1	0	M	0	−1	0	0.0000	0.0000
S	2	−2	2	A	−1	−2	2	0.0000	0.0000
R	0	−1	0	A	0	−1	0	0.0000	0.0000
S	1	1	1	A	2	1	2	1.4142	1.4142
G	2	−2	2	S	−1	−1	1	1.4142	2.8284
V	2	−2	2	V	−1	−2	2	0.0000	2.8284
A	0	−1	0	T	0	−1	0	0.0000	2.8284
V	2	−2	2	S	−1	−2	2	0.0000	2.8284
A	2	−2	2	P	−1	−2	2	0.0000	2.8284
D	−2	−2	−2	G	1	−2	−2	0.0000	2.8284
E	2	−2	2	S	−1	−1	1	1.4142	4.2426
S	−2	−2	−2	L	1	−2	−2	0.0000	4.2426
L	2	−2	2	E	−1	−2	2	0.0000	4.2426
T	0	0	0	L	−2	0	0	0.0000	4.2426
A	−2	−2	−2	L	1	−2	−2	0.0000	4.2426
F	0	−2	0	Q	0	−2	0	0.0000	4.2426
N	2	−1	2	P	−2	−1	2	0.0000	4.2426
D	2	−1	2	G	−2	−1	2	0.0000	4.2426
L	0	−2	0	F	0	−2	0	0.0000	4.2426
K	−2	−2	−2	S	1	−2	−2	0.0000	4.2426
L	2	−2	2	K	−1	1	−1	4.2426	8.4853
G	2	−2	2	T	−1	−1	1	1.4142	9.8995
K	−2	−2	−2	L	1	−2	−2	0.0000	9.8995
K	−1	−1	−1	L	0	−2	0	1.7321	11.6315
Y	0	2	0	G	0	2	0	0.0000	11.6315
K	2	−1	2	T	−2	−1	2	0.0000	11.6315
F	0	2	0	R	0	2	0	0.0000	11.6315
I	0	1	0	L	0	1	0	0.0000	11.6315
L	0	1	0	E	0	1	0	0.0000	11.6315
F	0	2	0	A	0	2	0	0.0000	11.6315

There we take the first 30 residues of the ND5 protein sequences of Human and common chimpanzee as an example. In Table 1, we showed the distance and the cumulative distance of every residue.

Figure 3 shows the distances of Human and common chimpanzee based on points $s_i = (x_i, y_i, z_i)$. As the figure shows, the zeros in the curve mean that at those points the Human and common chimpanzee are identical, otherwise, two sequences are different at those points.

Figure 4 shows the cumulative distances of Human and common chimpanzee based on points $s'_i = (X_i, Y_i, Z_i)$. In Figure 4, the diagonal parts of the curve mean that at those points the Human and common chimpanzee are different, while the horizontal parts mean that two sequences are identical.

Figure 5 shows the 2D curves based on X, Y, Z components of the first 30 residues of the ND5 protein sequences of Human and common chimpanzee, from which we can get that if the X_Human and X_common chimpanzee or Y_Human and Y_common chimpanzee or Z_Human and Z_common chimpanzee are not parallel, so the amino acids according to theses points are different.

As Figures 3, 4 and 5 show that, using our method, the differences of proteins can be found from the curves directly. There, we compute the cumulative distances of nine ND5 protein sequences (Randić *et al.*, 2009) as in Table 2.

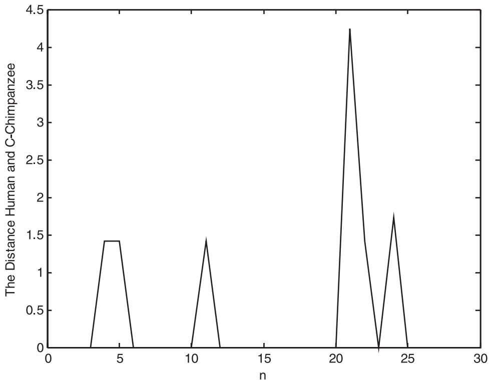

Fig. 3. The distances of human and common chimpanzee based on points $s_i = (x_i, y_i, z_i)$.

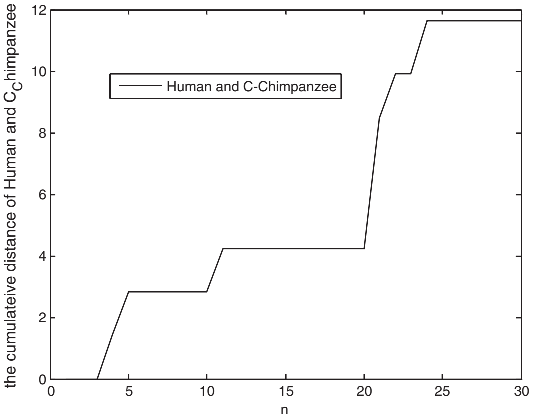

Fig. 4. The cumulative distances of human and common chimpanzee based on points $s'_i = (X_i, Y_i, Z_i)$.

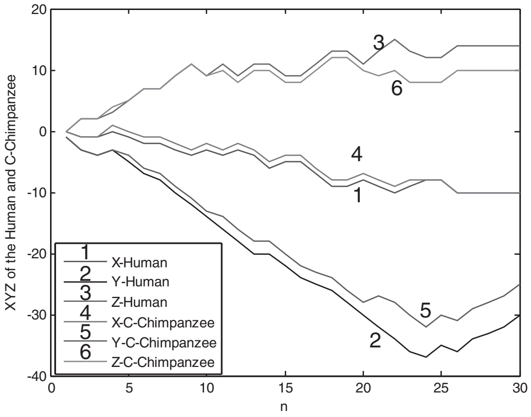

Fig. 5. The 2D curves based on X, Y, Z components of the first 30 residues of the ND5 protein sequences of human and common chimpanzee.

Table 2. The Information for Nine ND5 protein sequences

Species	ID	Length
Human	AP_000649	603
Gorilla	NP_008222	603
Pigmy chimpanzee	NP_008209	603
Common chimpanzee	NP_008196	603
Fin whale	NP_006899	606
Blue whale	NP_007066	606
Rat	AP_004902	610
Mouse	NP_904338	607
Opossum	NP_007105	602

Table 3. The cumulative distances for the nine ND5 protein sequences

	Gorilla	Pigmy chimpanzee	Common chimpanzee	Fin whal	Blue whal	Rat	Mouse	Opossum
Human	137.6	99.0	98.7	514.0	528.5	649.8	620.9	1014.8
Gorilla		123.2	127.6	527.5	541.1	648.0	611.4	1034.8
Pigmy chimpanzee			68.5	506.4	520.0	652.9	624.4	1014.6
Common chimpanzee				508.2	521.1	648.8	623.3	1022.0
Fin whal					53.0	613.6	605.3	1042.8
Blue whal						620.4	606.4	1054.4
Rat							379.8	1043.9
Mouse								1033.9

In Table 3, we list the cumulative distances among nine proteins. The smaller the distance means the two sequences are more similar. On the whole, we find that the proteins of fin whale–blue whale and common chimpanzee–pigmy chimpanzee are most similar; Human, Gorilla, common chimpanzee and pigmy chimpanzee are also very similar. Further more, ND5 protein of opossum is very dissimilar to others among the nine species. The results about the similarity are consistent to the known fact of evolution.

Cluster W is a multiple sequence alignment program for DNA or proteins. It produces biologically meaningful multiple sequence alignments of divergent sequences. It calculates the best match for the selected sequences, and lines them up so that the identities, similarities and differences can be seen. Evolutionary relationships can be seen via viewing Cladograms or Phylograms. So Cluster W is used to compute the similarity of sequences and construct the phylogenetic tree.

Compared with our Table 3 and He's Table 4 (He *et al.*, 2010), which was computed based on the Cluster W, we can find that the two tables are very similar. In the two tables, the smallest distance is between fin whale and blue whale, and the second smallest is between common chimpanzee and pigmy chimpanzee, followed by Human and common chimpanzee, human and pigmy chimpanzee, pigmy chimpanzee and Gorilla, common chimpanzee and Gorilla, Human and Gorilla, Rat and Mouse and so on.

Also, we construct a phylogenetic tree using fuzzy theory based on Table 3, which is shown in Figure 6. Immediately, we can find that our phylogenetic tree is consistent with He's Figure 7, which is constructed by Cluster W (He *et al.*, 2010). From the phylogenetic tree, we can find that the broken line between fin whale and blue whale is shortest, which means fin whale and blue whale are most similar. The second shortest broken line is between common chimpanzee and pigmy chimpanzee, followed by human, Gorilla,

Table 4. Comparison of prediction performance for different methods on the 98 apoptosis proteins dataset with a jackknife test

Algorithm	SEN (%)				ACC
	CY (43)	ME (30)	MI (13)	OTHER (12)	
Covariant[a]	97.7	73.3	30.8	25	72.5
BC[b]	90.7	90	92.3	50	85.7
Hens_BC[c]	95.3	90	92.3	66.7	89.8
ID_SVM[d]	95.3	93.3	84.6	58.3	88.8
LABSVM[e]	97.7	96.7	92.3	75	93.9
Our method	88.4	96.7	92.3	75	89.8

[a] The result based on the Covariant method (Zhou and Doctor, 2003).
[b] The result based on the BC method (Bulashevska and Eils, 2006).
[c] The result based on the Hens_BC method (Bulashevska and Eils, 2006).
[d] The result based on the ID_SVM method (Chen and Li, 2007a).
[e] The result based on the LABSVM method (Zhang *et al*., 2009).

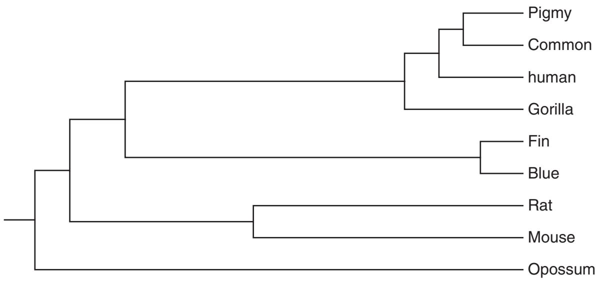

Fig. 6. Phylogenetic tree of the nine ND5 proteins based on our new graphical representation.

rat, mouse and Opossum, that is also the order of similar for the nine species.

3 A NEW METHOD FOR COMPUTING THE DISTANCE OF TWO SEQUENCES

There are some ways to analyze the similarity/dissimilarity of different sequences. The main is computing the distance; the smaller value of the distance represents the more similar of two sequences. The Euclidean distances between the leading Eigen values of some matrices was used, that matrices was constructed from the points of the sequences and was used to represent the sequences, which was named as mathematical descriptors, such as E, D/D, M/M, L/L, $^kL/^kL$, $^kM/^kM$, CM matrices (Cao *et al*., 2008; Liao and Zhu, 2006; Liao *et al*., 2005, 2007; Liu *et al*., 2009; Nandy, 1996; Randić *et al*., 2000a, b, 2003a, b; Yu and Sun, 2010; Yu *et al*., 2009). Further more, some researchers used the cumulative distance of

$$E\left(Sa,Sb\right)=\begin{cases}\dfrac{\sum\limits_{i=1}^{A}\sqrt{[Sa\,(i|x,y,z\ )-Sb\,(i|x,y,z\ )]^2}}{A} & \text{if } A=B\\[2ex] \dfrac{\sum\limits_{c=1}^{C}\sum\limits_{i=1}^{B}\sqrt{[Sa\,(c+i-1|x,y,z\ )-Sb\,(i|x,y,z\ )]^2}}{B\times C} & \text{if } A>B, C=A-B\end{cases} \tag{1}$$

every points or the last points distances (include the every component of points, or the angles distance) to be the sequences' distance (Randić *et al*., 2009).

$$[Sa(i|x,y,z)-Sb(i|x,y,z)]^2$$
$$=[x(Sa(i))-x(Sb(i))]^2+[y(Sa(i))-y(Sb(i))]^2+[z(Sa(i))-z(Sb(i))]^2$$

There we give a new distance computing method based on Euclidean distance. For example, two sequences, Sa, Sb; A, B are their lengths, respectively. If $A=B$, we compute the Euclidean distance directly. Presume $A>B$, computing the Euclidean distance of $Sa(1{:}B)$, $Sa(2{:}B+1)\ldots Sa(A-B+1:A)$ and Sb, respectively, then get the mean value of the $A-B$ distances to be the distances of the two sequences, as formula (1). This method can get more information in the sequences, so it is more effective.

Where $Sa(i|x, y, z)$ means the x, y, z components of the i-th point of sequence Sa.

4 THE APPLICATION OF OUR METHOD

Sub-cellular localization of proteins is closely related to proteins' function, while it is the protection of maintaining the highly ordered cell system operating normally. Studying the sub-cellular localization of proteins is helpful for understanding the nature and function of proteins, understanding the interaction between proteins and control mechanisms, and providing information for the invention of new drugs.

Biological cell is a highly ordered structure, based on the different distribution and function, which can be divided into different organelles, or cells areas, such as nucleus (NU), Golgi apparatus, endoplasmic reticulum (EN), mitochondria (MI), cytoplasm (CY) and cell membrane (ME) and so on. Synthesized proteins are transferred to specific organelles, for example, some of the proteins were transferred into the extracellular or remain in the cytoplasm. Only transferred to the correct location, proteins can be used in various life activities in cells. If the positioning error occurs, cell function or life will be impact badly. Proteins perform their biological functions just only in specific cell sites. The regional distribution of organelles can affect protein folding, aggregation and post-transcriptional modification process and make a great impact on cell function. Understanding of protein sub-cellular localization information can provide the necessary assistance for us to conclude the biological function of the protein, while provide the necessary information for other studies, such as protein interactions, evolution and so on.

There, we apply our method, the new representation and the distance got by the new method [based on s_i'$=(X_i$, Y_i, $Z_i)$], in the protein sub-cellular localization prediction by the similarity of sequences. We use the 98 apoptosis proteins (Zhang *et al*., 2009) dataset constructed by Zhou (Zhou and Doctor, 2003) and the 317 apoptosis proteins dataset constructed by Chen and Li (2007b). The 98 apoptosis proteins dataset contained 43 cytoplasm proteins (CY), 30 plasma membrane-bound proteins (ME), 13 mitochondrial inner and outer proteins (MI) and 12 other proteins (OTHER). The 317 dataset consisted of 317 apoptosis proteins divided into six sub-cellular locations with 112 cytoplasm proteins (CY), 55 membrane proteins (ME), 34 mitochondrial proteins (MI), 17 secreted proteins (SE) and 52 nuclear proteins (NU) and 47 endoplasmic reticulum proteins (EN). Dividing the sequences in the dataset into two parts, one to be sample set, the other is used as the test set. In the test set, one testing sequence can be classified as the class of the sequence

Table 5. Comparison of prediction performance for different methods on the 317 apoptosis proteins dataset with a jackknife test

Algorithm	SEN (%)						ACC
	CY (112)	ME (55)	MI (34)	SE (17)	NU (52)	EN (47)	
ID[a]	81.3	81.8	85.3	88.2	83	82.7	82.7
ID_SVM[b]	91.1	89.1	79.4	58.8	87.2	73.1	84.2
LABSVM[c]	92.9	85.5	76.5	76.5	93.6	86.5	88
Our method	88.4	85.5	76.5	58.8	78.9	91.5	83.6

[a] The result based on the ID method (Chen and Li, 2007b).

[b] The result based on the ID_SVM method (Chen and Li, 2007a).

[c] The result based on the LABSVM method (Zhang *et al*., 2009).

in the sample set which has the smallest distance with the testing sequence. We test the prediction accuracy in jackknife test.

The following measures were used to assess the performance of the classifiers used in this study: the over all prediction accuracy ACC, individual sensitivity SEN, individual specificity and Matthew's correlation coefficient MCC are used to measure the prediction performance of our work. The definition is showed as follows:

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{FN}+\text{FP}+\text{TN}} \tag{2}$$

$$\text{Sensitivity y} = \frac{\text{TP}}{\text{TP}+\text{FN}} \tag{3}$$

$$\text{Specificity y} = \frac{\text{TN}}{\text{FP}+\text{TN}} \tag{4}$$

$$\text{MCC} = \frac{(\text{TP}\times\text{TN})-(\text{FP}\times\text{FN})}{\sqrt{(\text{TP}+\text{FP})(\text{TP}+\text{FN})(\text{TN}+\text{FP})(\text{TN}+\text{FN})}} \tag{5}$$

where TP is the number of true positives, FP is the number of false positives, TN is the number of true negatives, FN is the number of false negatives and recall is equivalent to the sensitivity.

In Tables 4 and 5, we list our results for dataset 98 proteins and dataset 317 proteins in jackknife test, respectively. From Table 4, we can see that the overall accuracy for 98 proteins dataset by our method achieve 89.8%, which is higher than the covariant method (Zhou and Doctor, 2003), the BC method (Bulashevska and Eils, 2006), the ID_SVM method and equal to the Hens_BC (Bulashevska and Eils, 2006) method. Table 5 shows that the accuracy of our method is higher than the ID method. This result means our method is effective. But our method is lower than the LABSVM method (Zhang *et al*., 2009) in the 98 dataset and not as good as the ID_SVM (Chen and Li, 2007a) and the LABSVM (Zhang *et al*., 2009) method in the 317 dataset. But as we can see, our method did not use any classifier such as Bayesian classifier and SVM (Chen and Li, 2007a, b; Zhang *et al*., 2009) or other machine-learning classifiers, which can improve the accuracy by training and complicated computing, so our method is easier to carry out and less time consuming.

5 DISCUSSION

Our new graphical representation is seeking distribution property of amino acids from the nature of nucleotide triplets. As shown in Figure 2, the 2D distribution of the 20 amino acids is not unique. Our method is based on the principles of symmetry and the nearest. While using this method in phylogenetic tree construction, we select the sequences of only nine species. This nine species has been used in many other articles, including DNA phylogenetic tree construction and protein phylogenetic tree. Therefore, it has a strong representative, and biological evolution relations of this nine species are more clearly, really.

We proposed a new formula for computing the distance of sequences, which can deal with the sequences with unequal lengths. We know that mostly the lengths of sequences are not equal. Euclidean distance is easy to miss information or have additional information to add. Many of the existing distance calculations are obtained special matrixes from the graphical expression, and then raised invariants from the matrixes, at last, calculated distance of the invariants to represent the distance of sequences. Although those ways can handle the situation of unequal lengths too, they need to calculate the matrixes and invariants, which are complicated. Our method is proceeding directly from the graphic expression, to obtain the distance between sequences and can effectively deal with the situation of unequal length sequences.

In the sub-cellular localization, many articles used machine-learning methods, especially SVM method. These methods have to set up a training set, part of the dataset, to train the model and get the model parameters, and then use the model to test the testing set. Therefore, this method is computationally expensive, and complex. Our method uses the distance between sequences to measure the sequence difference, the smaller the difference, we think they tend to be more consistent with the location information. Although our results are not as well as some results of machine-learning methods, our way just only need to calculate the distance between the sequences, which is easy to operate.

6 CONCLUSION

In this article, we proposed a new graphical representation of proteins based on genetic-code distribution. We applied this method in the phylogenetic tree constructing, the result of which is consistent with the result got from a multiple sequence alignment program, Cluster W. Then we provided a new distance computing method based on the graphical representation and its application in the protein sub-cellular localization prediction using the similarity comparisons. The results show that our new graphical representation and new distance are more effective than some methods, although not as good as some machine-learning classifiers. But it is easier, because only need to compute the distance, not need machine learning or training.

However, as we can see, the result of our method is not the best, so in the future research; we would try to improve our method to increase the accuracy, and test in many different datasets, expecting our method will have general applicability.

Funding: National Nature Science Foundation of China (Grant 60973082, 60736016); Major State Basic Research Development Program of China (973 program Grant 2006CB303000); National Nature Science Foundation of Hunan province (Grant 07JJ5080, 06JJ2090); Planned Science and Technology Project of Hunan Province (Grant 2009FJ3195).

Conflict of Interest: none declared.

REFERENCES

Al-Shahib,A. *et al*. (2005a) Feature selection and the class imbalance problem in predicting protein function from sequence. *Appl. Bioinform.*, **4**, 195–203.

Al-Shahib,A. *et al*. (2005b) FRANKSUM: new feature selection method for protein function prediction. *Int. J. Neural Syst.*, **15**, 250–275.

Bai,F.L. and Wang,T.M. (2005) A 2-D graphical representation of protein sequences based on nucleotide triplet codons. *Chem. Phys. Lett.*, **413**, 458–462.

Bulashevska,A. and Eils,R. (2006) Predicting protein sub cellular locations using hierarchical ensemble of Bayesian classifiers based on Markov chains. *BMC Bioinformatics*, **7**, 298.

Chen,Y.L. and Li,Q.Z. (2007a) Prediction of apoptosis protein sub cellular location using improved hybrid approach and pseudo-amino acid composition. *J. Theor. Biol.*, **248**, 377–381.

Chen,Y.L. and Li,Q.Z. (2007b) Prediction of the sub cellular location of apoptosis proteins. *J. Theor. Biol.*, **245**, 775–783.

Cao,Z. *et al*. (2008) A group of 3D graphical representation of DNA sequences based on dual nucleotides. *Int. J. Quantum Chem.*, **108**, 1485–1490.

Gao,Q.B. *et al*. (2005) Prediction of protein sub cellular location using a combined feature of sequence. *Fed. Eur. Biochem. Soc.*, **579**, 3444–3448.

He,P.A. *et al*. (2010) The graphical representation of protein sequences based on the physicochemical properties and its applications. *J. Comput. Chem.*, **31**, 2136–2142.

Liao,B. *et al*. (2005) A 4D representation of DNA sequences and its application. *Chem. Phys. Lett.*, **402**, 380–383.

Liao,B. and Zhu,W. (2006) Analysis of similarity/dissimilarity of DNA primary sequences based on condensed matrices and information entropies. *Curr. Comput. Aid. Drug Des.*, **2**, 275–285.

Liao,B. *et al*. (2007) On the similarity of DNA primary sequences based on 5D representation. *J. Math. Chem.*, **42**, 47–57.

Lee,R.B. *et al*. (2009) Identification of protein functions using a machine-learning approach based on sequence-derived properties. *Proteome Sci.*, **7**, 27.

Li,X. *et al*. (2009) Protein functional class prediction using global encoding of amino acid sequence. *J. Theo. Biol.*, **261**, 290–293.

Liu,Z.B. *et al*. (2009) A 2-D graphical representation of DNA sequence based on dual nucleotides and its application. *Int. J. Quant. Chem.*, **109**, 948–958.

Nandy,A. (1996) Two-dimensional graphical representation of DNA sequences and intron-exon discrimination in intron-rich sequences. *Comput. Appl. Biosci.*, **12**, 55–62.

Randić,M. (2000a) Condensed Representation of DNA Primary Sequences. *J. Chem. Inform. Comput. Sci.*, **40**, 50–56.

Randić,M. *et al*.(2000b) On 3-D graphical representation of DNA primary sequences and their numerical characterization. *J. Chem. Inform. Comput. Sci.*, **40**, 1235–1244.

Randić,M. *et al*. (2003a) Novel 2-D graphical representation of DNA sequences and their numerical characterization. *Chem. Phys. Lett.*, **368**, 1–6.

Randić,M. *et al*. (2003b) Analysis of similarity/dissimilarity Of DNA sequences based on novel 2-D graphical representation. *Chem. Phys. Lett.*, **371**, 202–207.

Randić,M. *et al*. (2004) Unique graphical representation of protein sequences based on nucleotide triplet codons. *Chem. Phys. Lett.*, **397**, 247–252.

Randić,M. (2007) 2-D Graphical representation of proteins based on physico-chemical properties of amino acids. *Chem. Phys. Lett.*, **440**, 291–295.

Randić,M. *et al*. (2009) Graphical representation of proteins as four-color maps and their numerical characterization. *J. Mol. Graph. Model.*, **27**, 637–641.

Wen,J. and Zhang,Y.Y. (2009) A 2D graphical representation of protein sequence and its numerical characterization. *Chem. Phys. Lett.*, **476**, 281–286.

Yu,J.F. *et al*. (2009)TN curve: A novel 3D graphical representation of DNA sequence based on trinucleotides and its applications. *J. Theor. Biol.*, **261**, 459–468.

Yu,J.F. and Sun,X. (2010) Reannotation of protein-coding genes based on an improved graphical representation of DNA sequence. *J. Comput. Chem.*, **31**, 2126–2135.

Zhou,G P. and Doctor,K. (2003) Sub cellular location prediction of apoptosis proteins. *Proteins*, **50**, 44–48.

Zhang,L. *et al*.(2009) A novel representation for apoptosis protein subcellular localization prediction using support vector machine. *J. Theor. Biol.*, **259**, 361–365.