

# Protein Sequence Matching Using Parametric Spectral Estimate Scheme

Hsuan-Ting Chang\*, Hsiao-Wei Peng, Ciing-He Li  
Photonics and Information Laboratory  
Department of Electrical Engineering  
National Yunlin University of Science and Technology  
Douliu Yunlin

Neng-Wen Lo  
Department of Animal Science and Biotechnology  
Tunghai University  
Taichung, Taiwan

**Abstract**—Putative protein sequences decoded from the messenger ribonucleic acid (mRNA) sequences are composed of twenty amino acids with different physical-chemical properties, such as hydrophobicity and hydrophilicity (uncharged, positively charged or negatively charged amino acids). In this paper, the power spectral estimate (PSE) technique for random processes is applied to the protein sequence matching framework. First, the twenty kinds of amino acids are classified based on their hydrophobicity and hydrophilicity. Then each amino acid in the protein sequence is mapped to a corresponding complex value. Consider the various Hidden Markov chain orders in the complex valued sequences. The PSE method can explore the implicit statistical relations among protein sequences. The mean squared error between the power spectra of two sequences is determined and then used to measure their similarity. The experimental results verify that the proposed PSE method provides the consistent similarity measurement with the well-known ClustalW and BLASTp schemes. Moreover, the proposed PSE can show better similarity relevance than ClustalW and BLASTp schemes.

**Keywords**—protein sequence; amino acids; digital signal processing; parametric spectral estimate; hydrophilicity; hydrophobicity; Markov chain

## I. INTRODUCTION

In the past two decades, deoxyribonucleic acid (DNA) and protein sequences in various organisms have been massively obtained with the help of high-throughput sequencing technologies [1]. Biologists unravel the functionality and capability of numerous protein sequence domains by understanding their 3-D structures obtained by the x-ray diffraction technique or NMR technology. These procedures require laborious preparations of protein crystals and are extremely time-consuming. Therefore, alternative methods based on digital signal processing (DSP) technique were developed to circumvent the extremely complicated crystallographic tasks. Generally speaking, two types of methods are commonly used to analyze the protein sequences and predict their functions: (1) Statistical methods [2], which apply the well-known mathematical models in stochastic processes to analyze the sequences. (2) Geometrical methods [3], which apply graphs to represent the sequences and then analyze them. Both types of methods first transform the symbolic amino acids to numerical values. The global or local similarity of any two sequences can then be measured according to the differences between the extracted sequence

features. High similarity between two sequences may infer two meanings: (1) the two sequences could be homologous; (2) the protein structures and/or their biological functions are similar.

Recently, various methods are proposed to study the DNA and protein sequences. Among them are the DSP-based methods [4]-[7]. Some of the related studies put the focus on the visualization of sequences in various graphic forms [8]-[15]. In DSP techniques, each character in the DNA sequences or each amino acid in the protein sequence is mapped to a numerical value. According to the characteristics of the organisms, the different values used in the numerical mapping can be designed to accordingly fit the physical-chemical properties [16]-[21]. Thus, the comparison method especially utilizing certain properties of the residence in the DNA or protein sequence must be specifically designed. There are two well-known character-based tools for DNA and protein sequence comparison; ClustalW [22] and BLAST [23]. ClustalW is designed by using multiple-sequence alignment based on the meta-heuristics methods. The feature that the arrangement of each amino acid is similar in the evolution of the same species is utilized. On the other hand, BLAST alignment has four components: query, database, program, and search purpose/goal. BLAST is designed to locate the homologous sequence sites between two sequences using a heuristic approach. It compares partial sequences progressively, such that the local alignment results are obtained. Standard protein-protein BLAST (BLASTp) compares an amino acid query sequence against a protein sequence database. It is used for both identifying a query amino acid sequence and for finding similar sequences in protein databases.

In this paper, a parametric spectral estimate (PSE) method based on stochastic signal processing is proposed for protein sequence comparison. The numerical signals are used to represent the protein sequences and then analyzed in the frequency domain. First, a new model of mapping complex values to amino acids according to the physical-chemical characteristics is proposed. Next, the PSE method is used to determine the power spectrum density (PSD) of each numerical protein sequence. Finally, the mean squared error (MSE) values between two power spectra under various Markov orders are determined and served as a metric for sequence similarity measurement. As compared to the ClustalW and Blastp methods, our experimental results show that the proposed method provides an alternative way to efficiently

distinguish the differences between two protein sequences. The remaining part of this paper is organized as follows: Section 2 describes the proposed PSE method. The experimental results under different perspectives and their discussions are provided in Section 3. Finally, the conclusion is drawn in Section 4.

## II. METHODS

Figure 1 shows the block diagram of the proposed method for protein sequence comparison. There are three parts in this method: (1) numerical mapping; (2) PSE with different Markov orders; (3) MSE estimations, which are described in the following subsections.

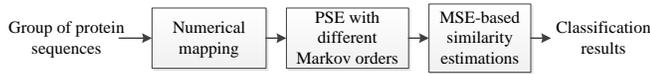


Fig. 1. The block diagram of the proposed method

### A. Numerical mapping

Table 1 shows the standard one-letter abbreviation of the 20 amino acids and their properties on charge, hydrophilicity or hydrophobicity, which are important for protein structure and protein-protein interaction. [24, 25].

According to the physical-chemical properties, twenty kinds of amino acids can be represented by twenty complex values which locate at the different positions on the unit circle in the complex plane. Figures 2(a)-2(d) show the four arrangements, Methods 1-4, of mapping the symbolic amino acids to the numerical ones. The twenty complex values are distributed on the unit circle whose center is at the origin of the complex plane. As shown in Fig. 2(a) (Method 1), the hydrophilic amino acids are distributed on the upper part of the circle. The first amino acid H is assigned to be  $45^\circ$  with respect to the real axis of the circle. The other amino acids {R, K, E, D, Q, N, Y, C, T, and S}, which have  $9^\circ$  separation from each other, are assigned after the amino acid H. The hydrophobic amino acids are distributed on the lower part of the circle. The first amino acid G is assigned to  $234^\circ$  with respect to the real axis on the circle. The other amino acids {A, V, L, I, P, M, F, and W}, which have  $9^\circ$  separation from each other, are assigned after the amino acid G. In addition to the hydrophobic and hydrophilic properties, we assign the positions according to their basic structures and the general chemical characteristics in their side chain (R) groups [26]. According to the position of the amino acids on the circle, the mapping is established so that every amino acid is adequately separated.

In addition to Method 1 shown in Fig. 2(a), three other mapping methods shown in Figs. 2(b)-2(d) are also proposed for performance comparison and evaluation. Table 2 shows the complex values corresponding to the coordinates of the 20 amino acids on the circle. In addition to Method 1, Methods 2, 3, and 4 are proposed to verify the effects of amino acid properties in this study by changing the positions on the circle. In Methods 2 and 3, the 20 amino acids still conform the rules of the characteristics of hydrophilicity and hydrophobicity, but the position can be exchanged in the random and horizontally reversed ways, respectively. In Method 4, the mappings are

random, and thus none of the rules of the characteristics of hydrophilicity and hydrophobicity is obeyed.

### B. Parameter spectrum estimation (PSE) method

Consider a stochastic process with the random variable (RV)  $\{X_n, n = 1, 2, 3, \dots, m\}$ , which describes a protein sequence composed of twenty kinds of amino acids and  $m$  is the sequence length. Let a sequence be denoted in Eq. (1), where  $i$  represents that the status is  $i$  when an amino acid is at the  $n$ th position in the sequence.

$$X_n = X_1, X_2, \dots, X_m, \quad X_n = i \in \{H, R, K, E, \dots, W\} \quad (1)$$

Let  $P_{ij}$  denote a transition probability given that the current status is  $i$  and the next status is  $j$ .

$$\begin{aligned} p_{ij} &= p\{X_n = j | X_{n-1} = i_{n-1}, X_{n-2} = i_{n-2}, \dots, X_2 = i_2, X_1 = i_1\} \\ &= p\{X_n = j | X_{n-1} = i_{n-1}\}, \quad \forall n > 1. \end{aligned} \quad (2)$$

Note that Eq. (2) denotes a Markov chain process, in which the probability of the current amino acid  $X_n$  is only dependent on the previous amino acid  $X_{n-1}$ . A one-step transition probability matrix of order  $20 \times 20$  is obtained by letting the first-order Markov chain model corresponding to the possible transitions between two amino acids in a protein sequence and is shown in Eq. (3).

$$P = [p_{ij}] = \begin{bmatrix} p_{HH} & p_{HR} & p_{HK} & \dots & p_{HW} \\ p_{RH} & p_{RR} & p_{RK} & \dots & p_{RW} \\ p_{KH} & p_{KR} & p_{KK} & \dots & p_{KW} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ p_{WH} & p_{WR} & p_{WK} & \dots & p_{WW} \end{bmatrix},$$

$$p_{ij} \geq 0, \sum_j p_{ij} = 1, i \in \{H, R, K, E, \dots, W\}. \quad (3)$$

Let the amino acids in a protein sequence be denoted as a discrete signal source where the occurrence probability of each element is  $p_i, i=1, 2, \dots, m$ . Equation (4) defines the information amount  $I(x_i)$  of  $x_i$  for an event which occurs with a probability  $p_i$ .

$$I(x_i) = \log_2\left(\frac{1}{p_i}\right), \quad \sum_{i=1}^m p_i = 1 \quad (4)$$

The average information or entropy of  $X$  is defined in Eq. (5):

$$H_2(X) = E\{I(x_i)\} = \sum_X p(x_i) \log_2\left(\frac{1}{p(x_i)}\right) \quad (5)$$

The conditional entropy of the current RV,  $X$ , given the  $m$  previous RVs:  $x_1, x_2, \dots, x_m$ , is defined in Eq. (6):

$$\begin{aligned} H_m(X | x_1, x_2, \dots, x_m) &= \\ \sum_X p(x_i | x_{i1}, x_{i2}, \dots, x_{im}) \log_2 \frac{1}{p(x_i | x_{i1}, x_{i2}, \dots, x_{im})} \end{aligned} \quad (6)$$

The entropy of a first-order Markov process is defined in Eq. (7):

$$H_1(X) = \sum_{x^2} p(x_{i1}, x_{i2}) \log_2 \left( \frac{1}{p(x_{i2} | x_{i1})} \right) \quad (7)$$

where  $p(x_{i1}, x_{i2})$  is the probability when the two values  $x_{i1}$  and  $x_{i2}$  occur together. The entropy of a second-order Markov process is defined in Eq. (8):

$$H_2(X) = \sum_{x^3} p(x_{i1}, x_{i2}, x_{i3}) \log_2 \left( \frac{1}{p(x_{i3} | x_{i2}, x_{i1})} \right) \quad (8)$$

The homologous gene sequences have similar entropies when a higher-order Markov process is used [27]. In a lower order, however, each set of homologous gene sequences have various entropy values. The  $q^{\text{th}}$ -order Markov model is shown in Eq. (9)

$$X[n] = \sum_{k=1}^q a_k X[n-k] + W[n] \quad (9)$$

where both RVs  $X$  and  $W$  are zero-mean and the variance  $\text{Var}\{W[n]\}$  of  $W$  is denoted as  $\sigma_W^2$ . In the estimation of random variables [28], the  $q$ -dimensional vector is denoted as  $X \triangleq (X[n-1], X[n-2], \dots, X[n-q])^T$ . The Equation

$E[Y | X] = \sum_{i=1}^N a_i X_i$  provides linear prediction estimate of the scalar random variable  $Y = X[n]$  in terms of the  $\mathbf{a}_k$ . The coefficients  $\mathbf{a} \triangleq (a_1, a_2, \dots, a_q)^T$  are determined as the solution of the orthogonal equation, and the optimum value of  $\mathbf{a}$ , denoted as  $\mathbf{a}_o$  is shown in Eq. (10).

$$\mathbf{a}_o^T = \mathbf{k}_{YX} \mathbf{K}_{XX}^{-1} \quad (10)$$

Equations (11) and (12) define the cross covariance vector and the covariance matrix of  $X[n]$ , respectively.

$$\mathbf{K}_{YX} = \begin{bmatrix} E\{X[n]X[n-1]\} \\ E\{X[n]X[n-2]\} \\ \vdots \\ E\{X[n]X[n-q]\} \end{bmatrix} = \begin{bmatrix} K_{XX}[1] \\ K_{XX}[2] \\ \vdots \\ K_{XX}[q] \end{bmatrix} \quad (11)$$

$$\begin{aligned} \mathbf{K}_{XX} &= \begin{bmatrix} E\{X[n-1]X[n-1]\} & E\{X[n-1]X[n-2]\} & \dots & E\{X[n-1]X[n-q]\} \\ E\{X[n-2]X[n-1]\} & E\{X[n-2]X[n-2]\} & \dots & E\{X[n-2]X[n-q]\} \\ & \vdots & & \\ E\{X[n-q]X[n-1]\} & E\{X[n-q]X[n-2]\} & \dots & E\{X[n-q]X[n-q]\} \end{bmatrix} \\ &= \begin{bmatrix} K_{XX}[0] & K_{XX}[1] & \dots & K_{XX}[q-1] \\ K_{XX}[-1] & K_{XX}[0] & \ddots & \vdots \\ \vdots & \ddots & \ddots & K_{XX}[1] \\ K_{XX}[-(q-1)] & \dots & K_{XX}[1] & K_{XX}[0] \end{bmatrix} \end{aligned} \quad (12)$$

To obtain a simple PSD estimate, the covariance function  $k_{XX}[\tau]$  is replaced as shown in [28], and the solution yields parameter estimates  $\hat{a}_1, \hat{a}_2, \dots, \hat{a}_q$ . Finally, the PSD estimate is defined in Eq. (13):

$$\hat{S}(\omega) = \frac{\hat{\sigma}_W^2}{|1 - \sum_{k=1}^q \hat{a}_k \exp(-j\omega k)|^2} \quad (13)$$

where  $\hat{\sigma}_W^2$  denotes the variance and  $\hat{a}_k$  is a parameter of RV. The covariance matching property when the PSD function is regarded to as an auto-regressive model is defined in Eq. (14).

$$\text{IFT}\{\hat{S}(\omega)\} = \hat{R}_x[\tau], \quad |\tau| \leq q', \quad (14)$$

where  $\text{IFT}\{\}$  denotes the inverse Fourier transform,  $\tau$  is the power spectral order,  $q'$  is the highest order, and  $\hat{R}_x$  is the auto-correlation function of  $\hat{S}(\omega)$ .

### C. MSE determination

When the PSD of each sequence in the  $q^{\text{th}}$  Markov order is determined by the use of PSE, the PSD values are normalized within the range [0, 1]. Next, the MSE defined in Eq. (15), is used to compare the similarity between two protein sequences  $y_i$  and  $y_j$  under the  $q^{\text{th}}$  Markov order.

$$\text{MSE}(y_i^q, y_j^q) = \sqrt{\frac{\sum_{k=1}^N [y_i^q(k) - y_j^q(k)]^2}{N}}, \quad (15)$$

where the parameter  $N$  is the length of protein sequence,  $y_i^q$  and  $y_j^q$  are the normalized PSD values under the  $q^{\text{th}}$  order transformed from one and the other protein sequences, respectively. If the lengths of two sequences are different, their PSDs are not of the same length, either. Thus the MSE cannot be directly computed. To solve this problem, two methods are used to make two sequences the same length. First, the shorter sequence is interpolated to be of the same length with the longer one and is denoted as  $y_{sI}$ . Second, the longer sequence is down-sampled to be of the same length with the shorter one and is denoted as  $y_{lS}$ . These two methods are determined by Bi-linear interpolation shown in Eqs. (16) and (17).

$$y_{sI}(j'_n) = \alpha_n \times y_{j+1} + (1 - \alpha_n) \times y_j, \quad j=1, 2, \dots, N_s, \quad (16)$$

$$\text{for } j'_n = \begin{cases} j_1, & \text{if } n=1 \\ \frac{N_s-1}{N_l-1} + j'_{n-1}, & \text{if } n>1 \end{cases}, \quad n=1, 2, \dots, N_l$$

$$y_{lS}(j'_n) = \alpha_n \times y_{j+1} + (1 - \alpha_n) \times y_j, \quad j=1, 2, \dots, N_l, \quad (17)$$

$$\text{for } j'_n = \begin{cases} \frac{N_l-1}{N_s+1} + j_1, & \text{if } n=1 \\ \frac{N_l-1}{N_s+1} + j'_{n-1}, & \text{if } n>1 \end{cases}, \quad n=1, 2, \dots, N_s$$

Here  $N_s$  and  $N_l$  are the lengths of the shorter and longer sequences, respectively,  $j'_n$  is between  $j$  and  $j+1$ , and  $\alpha_n = (j'_n - j)$ . In the proposed methods, the MSE based on

short sequences,  $MSE_s$ , and the MSE based on long sequences,  $MSE_l$ , under a certain Markov order  $q$ , are determined, respectively, to obtain the average value  $MSE_{final}$  shown in Eq. (18).

$$MSE_l = \sqrt{\frac{\sum_{j=1}^{N_l} [y_{sl}(j) - y_l(j)]^2}{N_l}}, \quad MSE_s = \sqrt{\frac{\sum_{j=1}^{N_s} [y_{ls}(j) - y_s(j)]^2}{N_s}},$$

$$MSE_{final} = \frac{MSE_l + MSE_s}{2}. \quad (18)$$

TABLE I. THE STANDARD ONE-LETTER ABBREVIATION OF THE 20 AMINO ACIDS AND THEIR PROPERTIES ON POLARITY, CHARGE, AND HYDROPHILICITY OR HYDROPHOBICITY

Amino acid	1-letter	Side-chain polarity	Side-chain	Hydrophobic index
Alanine	A	Nonpolar	Neutral	1.8
Arginine	R	Basic polar	Positive	-4.5
Asparagine	N	Polar	Neutral	-3.5
Aspartic acid	D	Acidic polar	Negative	-3.5
Cysteine	C	Nonpolar	Neutral	2.5
Glutamic acid	E	Acidic polar	Negative	-3.5
Glutamine	Q	Polar	Neutral	-3.5
Glycine	G	Nonpolar	Neutral	-0.4
Histidine	H	Basic polar	Positive (10%)	-3.2
Isoleucine	I	Nonpolar	Neutral	4.5
Leucine	L	Nonpolar	Neutral	3.8
Lysine	K	Basic polar	Positive	-3.9
Methionine	M	Nonpolar	Neutral	1.9
Phenylalanine	F	Nonpolar	Neutral	2.8
Proline	P	Nonpolar	Neutral	-1.6
Serine	S	Polar	Neutral	-0.8
Threonine	T	Polar	Neutral	-0.7
Tryptophan	W	Nonpolar	Neutral	-0.9
Tyrosine	Y	Polar	Neutral	-1.3
Valine	V	Nonpolar	Neutral	4.2

TABLE II. THE COMPLEX VALUES OF 20 AMINO ACIDS IN THE PROPOSED FOUR MAPPING METHODS

	Amino acids (property)	Method 1	Method 2	Method 3	Method 4
Hydrophilic amino acid	H (Basic)	0.707+i 0.707	0.309+i 0.951	-0.707+i 0.707	0.707+i 0.707
	R (Basic)	0.588+i 0.809	-0.707+i 0.707	-0.587+i 0.809	-0.588-i 0.809
	K (Basic)	0.454+i 0.891	0.156+i 0.987	-0.454+i 0.891	0.454+i 0.891
	E (Acidic and their Amide)	0.309+i 0.951	-0.309+i 0.951	-0.309+i 0.951	-0.309-i 0.951
	D (Acidic and their Amide)	0.156+i 0.987	0.454+i 0.891	-0.156+i 0.987	0.156+i 0.987
	Q (Acidic and their Amide)	i	-0.454+i 0.891	i	-i
	N (Acidic and their Amide)	-0.156+i 0.987	-0.156+i 0.987	0.156+i 0.987	-0.156+i 0.987
	Y (Aromatic)	-0.309+i 0.951	0.707+i 0.707	0.309+i 0.951	0.309-i 0.951
	C (Hydroxyl or Sulfur-containing)	-0.454+i 0.891	i	0.454+i 0.891	-0.454+i 0.891
	T (Hydroxyl or Sulfur-containing)	-0.587+i 0.809	-0.587+i 0.809	0.588+i 0.809	0.454-i 0.891
Hydrophobic amino acid	S (Hydroxyl or Sulfur-containing)	-0.707+i 0.707	0.588+i 0.809	0.707+i 0.707	-0.707+i 0.707
	G (Aliphatic)	-0.588-i 0.809	-i	0.588-i 0.809	0.588+i 0.809
	A (Aliphatic)	-0.454-i 0.891	-0.454-i 0.891	0.454-i 0.891	-0.454-i 0.891
	V (Aliphatic)	-0.309-i 0.951	0.156-i 0.987	0.309-i 0.951	0.309+i 0.951
	L (Aliphatic)	-0.156-i 0.987	-0.588-i 0.809	0.156-i 0.987	-0.156-i 0.987
	I (Aliphatic)	-i	-0.156-i 0.987	-i	i
	P (Cyclic)	0.156-i 0.987	-0.309-i 0.951	-0.156-i 0.987	0.156-i 0.987
	M (Hydroxyl or Sulfur-containing)	0.309-i 0.951	0.588-i 0.809	-0.309-i 0.951	-0.309+i 0.951
	F (Aromatic)	0.454-i 0.891	0.454-i 0.891	-0.454-i 0.891	-0.587+i 0.809
	W (Aromatic)	0.588-i 0.809	0.309-i 0.951	-0.588-i 0.809	0.588-i 0.809



shows the power spectra of the various Markov orders obtained by using the PSE method for the sequences in Group 1. The blue and red lines represent the normalized power spectral values of two sequences. The x-axis denotes the angular frequency ( $0 - 2\pi$ ), while the y-axis denotes the normalized PSD. For each pair of protein sequences, five power spectra corresponding to five Markov orders  $q=1-5$  are determined. If these two lines are close to each other in the same Markov order, we may infer that the two sequences are significantly related. For example, the PSDs in Figs. 3(a), 3(b), 3(f), 3(m), 3(n), and 3(o) are similar in each Markov order because all the sequence pairs are homologous protein sequences, while in Figs. 3(c), 3(d), 3(e), 3(g), 3(h), 3(i), 3(j), and 3(l) are obviously different because the sequence pairs are non-homologous.

Based on Methods 1-4 shown in Table 2, Tables 4 to 9 show the comparison results of the MSE values between various pairs of the sequences in the six groups, respectively. A smaller MSE value represents less difference between two sequences. In these six tables, all the MSE values of the homologous sequence pairs are smaller than 0.2. On the contrary, the MSE values of the non-homologous sequence pairs are greater than 0.2. This can be observed in the other parts in these tables as well. The proposed method contributes the classification of protein sequences and thus can serve as an alternative for the sequence comparison task.

Figures 4(a)-4(f) illustrate the comparison results of the proposed method, ClustalW, and BLASTp in Tables 4 to 9. The horizontal axis denotes the set of experiments (sequence pair), while the vertical axis denotes the normalized difference value. In order to compare the sequence similarly in each method accordingly, the BLASTp scores are replaced by the value 1-BLASTp to correspond to the same numerical characteristics with that in the proposed methods and ClustalW. In Table 3, some results of BLASTp method are shown as NF, which means that the two protein sequences have no similarity found. Here, we set the NF value as 0 and then the value of 1-BLASTp is 1, which denotes the maximal difference. In Fig. 4(a), the depicted lines of MSE1, MSE2, MSE3, MSE4, and ClustalW have the similar rising and descending trends for the short sequences. The scores in all the six methods basically can be used to distinguish the homologous and non-homologous sequences. However, in Figs. 4(b) and 4(c), MSE2 and MSE4 are not consistent to MSE1, ClustalW, and 1-BLASTp for the

medium and long sequence pairs. In Figs. 4(b) and 4(f), the 1-BLASTp scores are quite different from other scores and the variations among these methods are larger than that in the other figures. Note that the MSE4 values are higher than other values for Sequence pairs 14, and 15 in Figs. 4(d) to 4(f). The experimental results show that the mapping methods while encountering characteristics of hydrophilic amino acids and the general physical-chemical properties of amino acids can affect the comparison results. In Figs. 4(a) to 4(f), the MSE1 and MSE3 are nearly the same because the separation between each two amino acids is the same even if the mapping positions are horizontally reversed.

According to the results of the proposed methods, ClustalW, and BLASTp, the following observations are obtained: (1) The numerical mapping according to the physical-chemical characteristics of amino acids affect the results of comparison. In the experimental results, the more characteristics of protein are considered and arranged, the more correct results are obtained. (2) The MSE1 results are similar to the ClustalW scores, which means that the mapping scheme is consistent with the ClustalW method. (3) BLASTp and ClustalW are different methods, especially designed for the global and local sequence comparisons, respectively. The differences shown in the experimental results are especially obvious for the sequences in Groups 2 and 6.

#### IV. CONCLUSION

We proposed a new comparative tool for protein sequence comparison utilizing the parametric spectral estimate in stochastic processes to analyze protein sequences. The concepts of hydrophobicity in the amino acid physical-chemical properties are used to transform amino acids to numerical values. The experimental results show that the proposed methods effectively achieved the consistent comparison results with the well-known ClustalW and BLASTp. This research provides a new insight for the biologists as to how protein sequences can be analyzed. In our future work, more protein sequences will be tested by the proposed method. The problems encountered by two protein sequences with large difference in length will be tackled as well.

#### ACKNOWLEDGMENT

This research is partially supported by Ministry of Science and Technology (MOST), Taiwan under the contract number MOST 103-2221-E-224-035-MY2.

TABLE III. THE THREE GROUPS OF THE TEST PROTEIN SEQUENCES IN OUR EXPERIMENTS

	No.	Index	Class	Length (aa)
Group 1 (<200)	1	XP_933607.1	PREDICTED: hypothetical protein	90
	2	XP_001129788.1	PREDICTED: hypothetical protein	90
	3	XP_001129824.1	PREDICTED: hypothetical protein	90
	4	NP_001463.1	G antigen 2 [Homo sapiens]	116
	5	NP_001468.1	G antigen 7B [Homo sapiens]	117
	6	NP_036328.1	G antigen 8 [Homo sapiens]	116
Group 2 (300~400)	7	CAI46074.1	hypothetical protein [Homo sapiens]	329
	8	EFB16212.1	hypothetical protein PANDA_018121 [Ailuropoda melanoleuca]	329
	9	CAH89386.1	hypothetical protein [Pongo abelii]	323
	10	AAB51177.1	human RAD23A homolog [Homo sapiens]	363
	11	AAI33283.1	RAD23A protein [Bos taurus]	362
	12	AAH84695.1	RAD23 homolog A [Rattus norvegicus]	351
	13	AAH33781.1	PAXIP1 protein [Homo sapiens]	757

Group 3 (401~800)	14	CAD98066.1	hypothetical protein, partial [Homo sapiens]	675
	15	AAB91434.1	CAGF28, partial [Homo sapiens]	744
	16	AAP04006.1	NIMA-family kinase NEK8 [Homo sapiens]	692
	17	NP_001179652.1	serine/threonine-protein kinase Nek8 [Bos taurus]	698
	18	DAA19021.1	TPA: NIMA-related kinase 8-like [Bos taurus]	703
Group 4 (801~1200)	19	NP_031375.3	PAX-interacting protein 1 [Homo sapiens]	1069
	20	EHH17890.1	hypothetical protein EGK_14374, partial [Macaca mulatta]	1049
	21	AAH77588.1	K14 protein [Xenopus laevis]	1320
	22	CAC35387.1	suppression of tumorigenicity 5 [Homo sapiens]	1137
	23	AIC55177.1	ST5, partial [synthetic construct]	1137
	24	JAA49668.1	Putative ras signaling inhibitor st5 [Desmodus rotundus]	1137
Group 5 (1201~2000)	25	BAA13389.1	KIAA0259, partial [Homo sapiens]	1550
	26	AFJ70237.1	DNA topoisomerase 2-binding protein 1 [Macaca mulatta]	1527
	27	AAI51238.1	Topoisomerase (DNA) II binding protein 1 [Homo sapiens]	1435
	28	AAI12162.1	Tumor protein p53 binding protein 1 [Homo sapiens]	1972
	29	BAE06107.1	TP53BP1 variant protein, partial [Homo sapiens]	1984
	30	JAB19186.1	tumor suppressor p53-binding protein 1 isoform 1 [Callithrix jacchus]	1970
Group 6 (2001~3000)	31	BAA83718.1	RNA binding protein [Homo sapiens]	2752
	32	NP_057417.3	serine/arginine repetitive matrix protein 2 [Homo sapiens]	2752
	33	BAA20782.3	KIAA0324 protein [Homo sapiens]	2800
	34	AAK39635.1	DNA polymerase theta [Homo sapiens]	2724
	35	CAI56770.1	hypothetical protein [Homo sapiens]	2149
	36	ELK18077.1	DNA polymerase theta [Pteropus alecto]	2597

TABLE IV. THE COMPARISON RESULTS OF THE PROPOSED, CLUSTALW, AND BLASTP METHODS FOR GROUP 1 (<200 AA) SEQUENCES

Set of experiments	Sequence pairs	MSE <sub>1</sub>	MSE <sub>2</sub>	MSE <sub>3</sub>	MSE <sub>4</sub>	ClustalW	Blastp score	1-Blastp score
1	(1,2)	0.0919	0.0623	0.0919	0.0946	0.0111	0.98	0.02
2	(1,3)	0.0166	0.0348	0.0166	0.0182	0.0056	0.99	0.01
3	(1,4)	0.3502	0.2343	0.3504	0.2682	0.4667	NF	1
4	(1,5)	0.3472	0.2273	0.3475	0.2545	0.4667	NF	1
5	(1,6)	0.3525	0.2346	0.3527	0.2605	0.4667	NF	1
6	(2,3)	0.0864	0.0572	0.0864	0.1043	0.0056	0.99	0.01
7	(2,4)	0.3755	0.2532	0.3751	0.2887	0.4667	NF	1
8	(2,5)	0.3718	0.2448	0.3715	0.2714	0.4667	NF	1
9	(2,6)	0.3770	0.2537	0.3767	0.2815	0.4611	NF	1
10	(3,4)	0.3540	0.2414	0.3540	0.2697	0.4667	NF	1
11	(3,5)	0.3505	0.2332	0.3507	0.2580	0.4667	NF	1
12	(3,6)	0.3564	0.2417	0.3564	0.2613	0.4667	NF	1
13	(4,5)	0.0567	0.0684	0.0567	0.0504	0.0086	0.97	0.03
14	(1,2)	0.0171	0.0055	0.0171	0.0247	0.0043	0.99	0.01
15	(1,3)	0.0585	0.0653	0.0586	0.0441	0.0129	0.97	0.03

TABLE V. THE COMPARISON RESULTS OF THE PROPOSED, CLUSTALW, AND BLASTP METHODS FOR GROUP 2 (300 – 400 AA) SEQUENCES

Set of experiments	Sequence pairs	MSE <sub>1</sub>	MSE <sub>2</sub>	MSE <sub>3</sub>	MSE <sub>4</sub>	ClustalW score	BLASTp score	1- BLASTp score
1	(7,8)	0.0409	0.1644	0.0409	0.0719	0.0304	0.94	0.06
2	(7,9)	0.1331	0.1834	0.1331	0.2182	0.2307	0.54	0.46
3	(7,10)	0.3324	0.3035	0.3325	0.1600	0.4863	0.71	0.29
4	(7,11)	0.3119	0.3181	0.3120	0.1860	0.4863	0.71	0.29
5	(7,12)	0.2822	0.2702	0.2823	0.1572	0.4742	0.83	0.17
6	(8,9)	0.1383	0.2212	0.1383	0.2146	0.2260	0.54	0.46
7	(8,10)	0.3309	0.3022	0.3312	0.1467	0.4620	0.71	0.29
8	(8,11)	0.3129	0.3130	0.3132	0.1736	0.4666	0.71	0.29
9	(8,12)	0.2862	0.2926	0.2864	0.1375	0.4681	0.71	0.29
10	(9,10)	0.3268	0.2768	0.3270	0.2615	0.4598	0.41	0.59
11	(9,11)	0.3183	0.3027	0.3186	0.2810	0.4721	0.41	0.59
12	(9,12)	0.2835	0.2320	0.2838	0.2584	0.4861	0.30	0.70
13	(10,11)	0.1175	0.2194	0.1175	0.0466	0.0221	0.93	0.07
14	(10,12)	0.1957	0.2301	0.1958	0.0421	0.0627	0.94	0.06
15	(11,12)	0.1653	0.2967	0.1654	0.0598	0.0698	0.91	0.09

TABLE VI. THE COMPARISON RESULTS OF THE PROPOSED, CLUSTALW, AND BLASTP METHODS FOR GROUP 3 (401 – 800 AA) SEQUENCES

Set of experiments	Sequence pairs	MSE1	MSE2	MSE3	MSE4	ClustalW score	BLASTp score	1-BLASTp score
1	(13,14)	0.0823	0.0272	0.0823	0.0330	0.0007	0.99	0.01
2	(13,15)	0.1413	0.2633	0.1413	0.3128	0.0894	0.96	0.04
3	(13,16)	0.3831	0.4798	0.3832	0.2793	0.4523	0.28	0.72
4	(13,17)	0.3687	0.4880	0.3688	0.2105	0.4592	0.27	0.73
5	(13,18)	0.3770	0.4945	0.3771	0.2209	0.4595	0.28	0.72
6	(14,15)	0.1570	0.2747	0.1569	0.3338	0.1089	0.96	0.04
7	(14,16)	0.3769	0.4823	0.3769	0.2879	0.4511	0.28	0.72
8	(14,17)	0.3666	0.4910	0.3666	0.2097	0.4578	0.28	0.72
9	(14,18)	0.3702	0.4976	0.3720	0.2203	0.4578	0.28	0.72
10	(15,16)	0.4245	0.4603	0.4246	0.3161	0.4798	0.44	0.56
11	(15,17)	0.4021	0.4617	0.4022	0.3442	0.4807	0.33	0.67
12	(15,18)	0.4044	0.4676	0.4044	0.3478	0.4808	0.33	0.67
13	(16,17)	0.0702	0.0816	0.0703	0.1614	0.0260	0.94	0.06
14	(16,18)	0.0750	0.0798	0.0751	0.1547	0.0318	0.94	0.06
15	(17,18)	0.0264	0.0214	0.0264	0.0279	0.0064	1.00	0

TABLE VII. THE COMPARISON RESULTS OF THE PROPOSED, CLUSTALW, AND BLASTP METHODS FOR GROUP 4 (801 – 1200 AA) SEQUENCES

Set of experiments	Sequence pairs	MSE <sub>1</sub>	MSE <sub>2</sub>	MSE <sub>3</sub>	MSE <sub>4</sub>	ClustalW score	BLASTp score	1- BLASTp score
1	(19,20)	0.1147	0.1193	0.1148	0.0734	0.0186	0.96	0.04
2	(19,21)	0.1811	0.1667	0.1814	0.1116	0.1240	0.84	0.16
3	(19,22)	0.3297	0.3615	0.3298	0.3311	0.4528	0.16	0.84
4	(19,23)	0.3326	0.3628	0.3327	0.3313	0.4511	0.16	0.84
5	(19,24)	0.3586	0.3763	0.3587	0.3442	0.4602	0.29	0.71
6	(20,21)	0.1925	0.1113	0.1928	0.0605	0.1239	0.82	0.18
7	(20,22)	0.3396	0.3470	0.3397	0.3527	0.4561	0.40	0.60
8	(20,23)	0.3452	0.3498	0.3452	0.3518	0.4557	0.40	0.60
9	(20,24)	0.3677	0.3710	0.3677	0.3787	0.4576	0.29	0.71
10	(21,22)	0.3011	0.3714	0.3011	0.3738	0.4705	0.37	0.63
11	(21,23)	0.3154	0.3738	0.3154	0.3721	0.4705	0.30	0.70
12	(21,24)	0.3266	0.3821	0.3265	0.4029	0.4776	0.26	0.74
13	(22,23)	0.0319	0.0235	0.0319	0.0319	0.0022	0.99	0.01
14	(22,24)	0.0813	0.1339	0.0813	0.1763	0.0299	0.94	0.06
15	(23,24)	0.0881	0.1329	0.0881	0.1842	0.0312	0.94	0.06

TABLE VIII. THE COMPARISON RESULTS OF THE PROPOSED, CLUSTALW, AND BLASTP METHODS FOR GROUP 5 (1201 – 2000 AA) SEQUENCES

Set of experiments	Sequence pairs	MSE <sub>1</sub>	MSE <sub>2</sub>	MSE <sub>3</sub>	MSE <sub>4</sub>	ClustalW score	BLASTp score	1- BLASTp score
1	(25,26)	0.0298	0.0390	0.0298	0.0811	0.0111	0.98	0.02
2	(25,27)	0.0377	0.0456	0.0376	0.0716	0.0000	1.00	0.00
3	(25,28)	0.3208	0.2858	0.3208	0.4020	0.4452	0.30	0.70
4	(25,29)	0.3238	0.2841	0.3238	0.4076	0.4445	0.30	0.70
5	(25,30)	0.2909	0.2970	0.2910	0.4384	0.4287	0.28	0.72
6	(26,27)	0.0594	0.0562	0.0594	0.1174	0.0098	0.98	0.02
7	(26,28)	0.3218	0.2835	0.3218	0.3992	0.4601	0.23	0.77
8	(26,29)	0.3245	0.2829	0.3245	0.4070	0.4430	0.23	0.77
9	(26,30)	0.2892	0.2974	0.2892	0.4289	0.4280	0.26	0.74
10	(27,28)	0.3166	0.2866	0.3166	0.3912	0.4408	0.33	0.67
11	(27,29)	0.3198	0.2868	0.3198	0.3970	0.4401	0.33	0.67
12	(27,30)	0.2905	0.3018	0.2906	0.4190	0.4300	0.26	0.74
13	(28,29)	0.0138	0.0120	0.0138	0.0325	0.0005	0.99	0.01
14	(28,30)	0.0922	0.0694	0.0922	0.1595	0.0251	0.95	0.05
15	(29,30)	0.0877	0.0698	0.0877	0.1715	0.0244	0.95	0.05

TABLE IX. THE COMPARISON RESULTS OF THE PROPOSED, CLUSTALW, AND BLASTP METHODS FOR GROUP 6 (2001 – 3000 AA) SEQUENCES

Set of experiments	Sequence pairs	MSE <sub>1</sub>	MSE <sub>2</sub>	MSE <sub>3</sub>	MSE <sub>4</sub>	ClustalW score	BLASTp score	1- BLASTp score
1	(31,32)	0.0026	0.0026	0.0026	0.0012	0.0002	0.99	0.01
2	(31,33)	0.0059	0.0094	0.0059	0.0023	0.0004	0.99	0.01
3	(31,34)	0.3391	0.4435	0.3391	0.5081	0.4670	0.88	0.12
4	(31,35)	0.3242	0.4118	0.3241	0.5117	0.4591	0.88	0.12
5	(31,36)	0.3251	0.3371	0.3251	0.4379	0.4704	0.44	0.56
6	(32,33)	0.0061	0.0077	0.0061	0.0023	0.0002	0.99	0.01
7	(32,34)	0.3384	0.4443	0.3384	0.5079	0.4670	0.88	0.12
8	(32,35)	0.3233	0.4127	0.3233	0.5115	0.4591	0.88	0.12
9	(32,36)	0.3245	0.3375	0.3245	0.4378	0.4704	0.44	0.56
10	(33,34)	0.3391	0.4485	0.3391	0.5078	0.4662	0.88	0.12
11	(33,35)	0.3241	0.4172	0.3240	0.5114	0.4581	0.88	0.12
12	(33,36)	0.3254	0.3385	0.3254	0.4373	0.4704	0.44	0.56
13	(34,35)	0.0998	0.1003	0.0999	0.1355	0.0016	0.99	0.01
14	(34,36)	0.1142	0.254	0.1142	0.3358	0.09261	0.81	0.19
15	(35,36)	0.1394	0.2791	0.1394	0.3236	0.10051	0.80	0.20

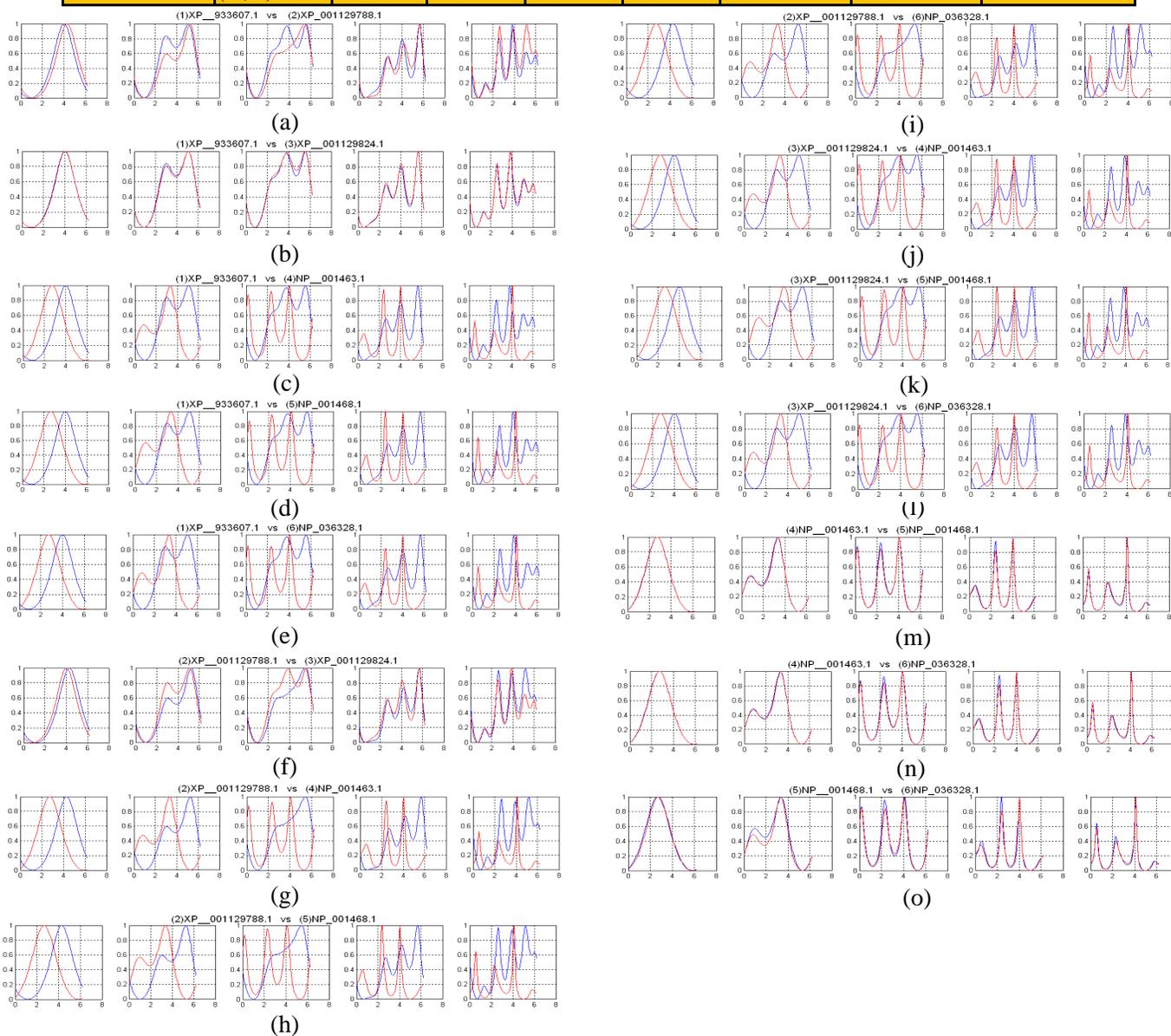


Fig. 3. The PSE comparison results between each sequence pair in Group 1 under various Markov orders  $q=1-5$

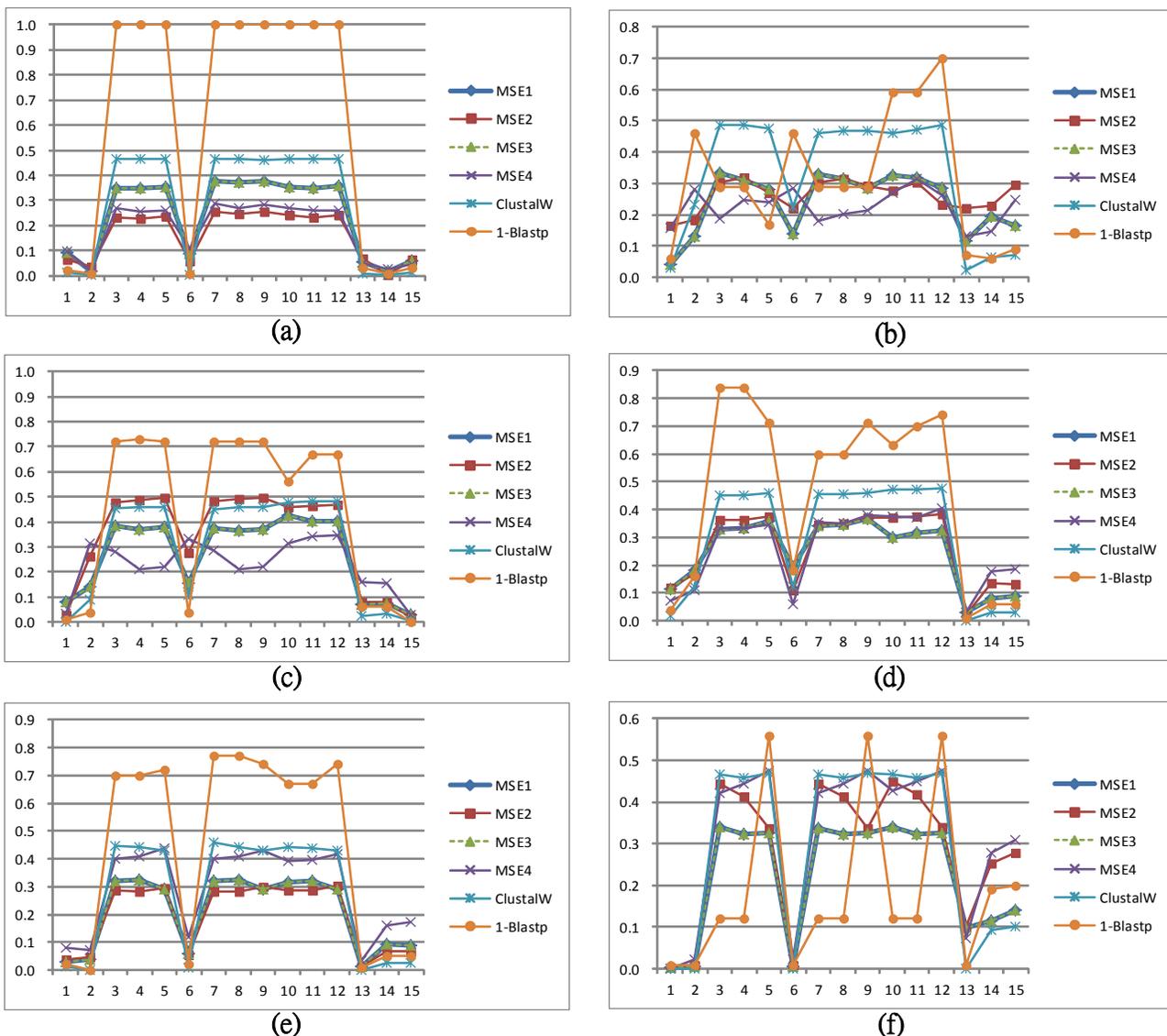


Fig. 4. The graphical representations of the comparison results on MSE1, MSE2, MSE3, MSE4, ClustalW, and 1-BLASTp in (a) Table 3; (b) Table 4; (c) Table 5; (d) Table 6; (e) Table 7; and (f) Table 8

REFERENCES

- [1] M. Guarnaccia, G. Gentile, E. Alessi, C. Schneider, S. Petralia, and S. Cavallaro, "Is this the real time for genomics?" *Genomics*, vol. 103, no. 2-3, pp. 177-182, 2014.
- [2] G. A. Price, G. E. Crooks, R. E. Green, and S. E. Brenner, "Statistical evaluation of pairwise protein sequence comparison with the Bayesian bootstrap," *Bioinformatics*, vol. 21, pp. 3824-3831, 2005.
- [3] Z. G. Yu, V. Anh, and K. S. Lau, "Chaos game representation of protein sequence based on the detailed HP model and their multifractal and correlation analyses," *Journal of Theoretical Biology*, pp. 341-348, 2004.
- [4] E. A. Cheever, D. B. Searls, W. Karunaratne, and G. C. Overton, "Using signal processing techniques for DNA sequence comparison," *Proceedings of the 1989 Fifteenth Annual Northeast Bioengineering Conference*, pp. 173-174, 1989.
- [5] D. Anastassiou, "DSP in genomics: processing and frequency-domain analysis of character strings," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, pp. 1053-1056, 2001.
- [6] P. Cristea, "Genetic signal analysis," *Sixth International Symposium on Signal Processing and Its Applications (ISSA 2001)*, vol. 2, pp. 703-706, 2001.
- [7] H. T. Chang, C.J. Kuo, N.-W. Lo, and W.-Z. Lv, "DNA sequence visualization and comparison based on quaternion number system," *International Journal of Advanced Computer Science and Applications*, vol. 3, no. 11, pp. 39-46, 2012.
- [8] D. Wu, J. Roberge, D. J. Cork, B. G. Nguyen, and T. Grace, "Computer visualization of long genomic sequence," *Proceedings of IEEE Conferences on Visualization*, pp. 308-315, 1993.
- [9] E. H.-H Chi, P. Barry, E. Shoop, J. V. Carlis, E. Retzel, and J. Riedl, "Visualization of biological sequence similarity search results," *Proceedings of IEEE Conferences on Visualization*, pp. 44-51, 1995.
- [10] H. H. Chi, J. Riedl, E. Shoop, J. V. Carlis, E. Retzel, and P. Barry, "Flexible information visualization of multivariate data from biological sequence similarity searches," *Proceedings of IEEE Conferences on Visualization*, pp. 133-140, 1996.
- [11] H.T. Chang, N.W. Lo, W.C. Lu, and C.J. Kuo, "Visualization and comparison of DNA sequences by use of three-dimensional trajectory," *The First Asia-Pacific Bioinformatics Conference (APBC 2003)*, vol. 19, pp. 81-85, Adelaide, Australia 2003.
- [12] H.T. Chang, S. W. Xiao, and N.-W. Lo, "Feature extraction and comparison of TDTs: an efficient sequence retrieval algorithm for

- genomic databases,” The Third Asia-Pacific Bioinformatics Conference (APBC 2005), pp. 86, 2005.
- [13] N.-W. Lo, H.T. Chang, S.W. Xiao, and C.J. Kuo, “Global visualization of DNA sequences by use of three-dimensional trajectories,” *Journal of Information Science and Engineering*, vol. 23, no. 6, pp. 1723-1736, Nov. 2007.
- [14] H.T. Chang, “DNA sequence visualization,” in *Advanced Data Mining Technologies in Bioinformatics*, Chapter 4, pp. 63-84, Edited by Dr. Hui-Huang Hsu, Idea Group Publishing, ISBN 195140864-4, 2006.
- [15] H.T. Chang, S.W. Xiao, and C.H. Lee, “Feature extraction for fast data retrieval for visualized DNA sequence,” *National Symposium on Telecommunications*, Kaoshiung Taiwan, 2006.
- [16] W. Wong and D. H. Johnson, “Computing linear transforms of amino acidic signals,” *IEEE Transactions on Signal processing*, vol. 50, pp. 628-634, 2002.
- [17] P. D. Cristea, “Conversion of nucleotides sequences into genomic signals,” *Journal of Cellular and Molecular Medicine*, vol. 6, no 2, pp. 279-303, 2002.
- [18] P.J.S.G. Ferreira, V. Afreixo, and D. Santos, “Spectrum and amino acid distribution of nucleotide sequences,” *Digital Signal Processing*, vol. 14, pp. 523-530, 2004.
- [19] R. Gupta, A. Mittal, and S. Gupta, “An efficient algorithm to detect palindromes in DNA sequences using periodicity transform,” *Signal Processing*, vol. 86, pp. 2067-2073, 2006.
- [20] X. Meng and V. Chaudhary, “A high-performance heterogeneous computing platform for biological sequence analysis,” *IEEE Transactions on Parallel and Distributed Systems*, vol. 21, pp. 1267-1280, 2010.
- [21] N. Maillat, C. Lemaitre, R. Chikhi, D. Lavenier, and P. Peterlongo, “Compareads: comparing huge metagenomic experiments,” *BMC Bioinformatics*, vol. 13, (Suppl 19): S10, 2012.
- [22] <http://www.clustal.org/clustal2/>
- [23] <http://blast.ncbi.nlm.nih.gov/>
- [24] R. E. Hausman and B. M. Cooper, *The cell: a molecular approach*. Washington, D.C: ASM Press. pp. 51. ISBN 0-87893-214-3, 2004.
- [25] J. Kyte and R. Doolittle, “A simple method for displaying the hydropathic character of a protein,” *Journal of Molecular Biology*, vol. 157, no. 1, pp. 105–32, May 1982
- [26] <https://www.boundless.com/biology/definition/r-group/> (R group)
- [27] H.T. Chang, Y.-L. Liu, C.J. Kuo, C.H. Lee, and N.-W. Lo, “Analysis of genomic DNA sequence by use of parametric spectral estimate,” *Bioinformatics in Taiwan (BIT2005)*, pp. 68, Tainan Taiwan, Sep. 2005
- [28] H. Stark and J. W. Woods, *Probability and Random Process with Applications to Signal Processing*, Chapter 9. Applications to Statistical Signal Processing, pp. 553-619, Third Edition, Prentice Hall, ISBN-10: 0130200719, 2001.