A Fractal Dimension and Wavelet Transform Based Method for Protein Sequence Similarity Analysis

Lina Yang, Yuan Yan Tang, Yang Lu, and Huiwu Luo

Abstract—One of the key tasks related to proteins is the similarity comparison of protein sequences in the area of bioinformatics and molecular biology, which helps the prediction and classification of protein structure and function. It is a significant and open issue to find similar proteins from a large scale of protein database efficiently. This paper presents a new distance based protein similarity analysis using a new encoding method of protein sequence which is based on fractal dimension. The protein sequences are first represented into the 1-dimensional feature vectors by their biochemical quantities. A series of Hybrid method involving discrete Wavelet transform, Fractal dimension calculation (HWF) with sliding window are then applied to form the feature vector. At last, through the similarity calculation, we can obtain the distance matrix, by which, the phylogenic tree can be constructed. We apply this approach by analyzing the ND5 (NADH dehydrogenase subunit 5) protein cluster data set. The experimental results show that the proposed model is more accurate than the existing ones such as Su's model, Zhang's model, Yao's model and MEGA software, and it is consistent with some known biological facts.

Index Terms—Protein sequence similarity, fractal dimension, Higuchi's algorithm, discrete wavelet transform, sliding window

1 Introduction

Proteins are constructed from 20 amino acids (referred to as residues) and are important parts of most biological processes. The rapid growth of the number of protein sequences has created many challenges for biologists, for which increasing efforts have been made to find the proper and reliable methods to analyze the vast amount of protein sequence data. The key techniques related to protein sequence analysis are to find the similar segments, domains between structures and functions or sequences which behaves similarly in biological process [1], [2], [3], [4]. Rigden [5] presented that proteins with significant similar sequence are likely to have similar function. These similarities help to identify individual protein, which is crucial for the biological research of their relation between their functions and structures [6], since identifying the protein function is a challenging task. It can also provide a better understanding of protein evolution which aids the identification of protein function and could lead to advances in biology as well as new treatments or drug developing for diseases.

Most of the existing mathematical approaches for protein sequences comparison are based on sequence alignments [7], which have some defects. For example, PSI-BLAST [8] and methods based on hidden Markov models (HMM) [9] have been developed for classifying protein sequences by searching against protein databases by their sequence alignments. However, these methods are to build a model for a single protein family and evaluate the fitness of each candidate sequence in the model [4]. They will not work when query proteins are lack of significant sequence similarity against the database sequences [10]. One of the popular models is the edit distance [11]. It concentrates on the minimum edit operations, e.g. insertion, deletion, or substitution when transforming one sequence to another. However, it involves in few biochemical features of protein. Some approaches divided protein sequence into different segments [12]. However, the optimal length of the small peptide fragments is hard to know or time-consuming to explore during division. Besides, Cheung and Jia [2] presented a general clustering framework based on the concept of object-cluster similarity and gives a unified similarity metric which can be simply applied to the data with categorical, numerical, and mixed attributes. Tsuda et al. [3] proposed an efficient method of protein classification using multiple protein networks. It assigned weights to multiple networks, and thereby select important ones. Huang, et al. [4] presented a new method for classifying protein sequences based on the hydropathy blocks occurring in protein sequences. A fixed-dimensional feature vector is first generated for each protein sequence using the frequency of the hydropathy blocks occurring in the sequence.

As an effective tool of signal processing, wavelet transform (WT) is widely used in bioinformatics and chemometrics and it shows the advantage in analyzing different scales of information of signals and bioinformatical data [13]. A survey of how wavelet analysis has been applied to cancer bioinformatics questions is provided in [14]. WT has been applied to predict the hydrophobic cores from hydropathy data [15]. It has also been used to predict the location and topology of helices in transmembrane proteins to predict protein secondary

- *The authors are with the Department of Computer and Information Science, Faculty of Science and Technology, University of Macau, 999078 Macau. E-mail: {yb27411, yytang, mb25426, yb17409}@umac.mo.*

Manuscript received 18 Feb. 2014; revised 5 Oct. 2014; accepted 9 Oct. 2014. Date of publication 15 Oct. 2014; date of current version 3 Apr. 2015.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TCBB.2014.2363480

1545-5963 © 2014 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

structures [16]. Based on discrete wavelet transform (DWT), a new concept of similarity of protein sequence, sequence-scale similarity, has been proposed in [17] to identify the functional similarity of two protein sequences. Besides, DWT was also applied with various protein substitution model to measure the pair-wise similarity [18]. A compound method for protein secondary structure prediction was proposed in [19]. The residue sequence was first substituted with corresponding hydrophobic values, then the sequential hydrophobic values was processed by using DWT for denoising. DWT was used to the pre-processing of spectra corresponding to several brain tumor pathologies [20].

Fractal dimension (FD) contains information about the geometrical structure. It is a useful concept in describing natural objects, which gives their degree of complexity [21]. Fractal geometry has been applied to the cardiovascular morphology problem. The authors have shown that the His-Purkinje system can be represented as a fractal network [22]. Boxt et al. illustrate that the fractal dimension of the pulmonary arterial tree is reduced of the development of hyperoxic pulmonary arterial hypertension in rats [23]. Chen et al. apply the theory of fractal dimension to analyze the similarity of the protein's structural spectrum, thus, the protein's spacial structure could be compared with both the global structural and the local structural [24]. Tao et al. propose a computational methodology based on fractal geometry for determining 3D structure of protein with imagery projection operations [25]. Holden et al. apply the fractal dimension and Shannon entropy to analyze the nucleotide fluctuation of the glycoprotein and nucleoprotein sequences in the Nipah virus [26]. Zhang and Kinsner present a multifractal techniques to estimate the multifractal measures of DNA sequences [27].

Although the fractal dimension has such wide applications, there is nobody applying it to protein sequence comparison in bioinformatics yet. This paper aims at exploring the fractal dimension calculation and DWT in protein sequence analysis for similarity comparison. It is anticipated that this approach can be widely used to process various types of protein sequence.

The rest of this paper is organized as follows. Section 2 describes the basic concepts and theories of fractal dimension. Our proposed algorithm will be discussed in Section 3. Section 4 presents the experimental results by comparing our proposed method with other methods. The conclusion is drawn in Section 5.

2 Basic Concepts of Fractal Dimension

This section introduces the basic theoretical framework of several kinds of fractal dimensions as well as several approximation algorithms for fractal dimension calculation which will be used in our proposed algorithm.

2.1 Theoretical Fractal Dimension

The term FD refers to a non-integer dimension of any object. FD analysis is frequently used in biomedical signal processing [28], image segmentation [29], analysis of audio signals [30]. In this section, the basic concepts and properties of fractal dimension will be introduced, and several types of fractal dimension such as Hausdorff dimension [31], box counting dimension, and Minkowski dimension [32] will also be discussed.

2.1.1 Hausdorff Dimension

Many kind of fractal dimensions are proposed by mathematicians, and among them the most important ones is Hausdorff dimension. It has the advantage of being defined for any set because of its mathematical convenience. The basic concept of Hausdorff dimension [31] will be presented below:

Assume U be a non-empty subset of n-dimensional Euclidean space $\mathbb{R}^n$, and the diameter of U is defined as

$$|U| = sup\{|x-y| : x, y \in U\},$$

where $sup\{.\}$ stands for the supremum of $\{.\}$. Thus, the diameter of U is the longest distance between any pair of points in U. If $\{U_i\}$ is a countable collection of sets of diameter at most ϵ that cover S, such that

$$S \subset \bigcup_{i=1}^{\infty} U_i, \qquad 0 < |U_i| \leq \epsilon,$$

we say that $\{U_i\}$ is a ϵ-cover of S.

Suppose that $S \subset \mathbb{R}^n$ and d is a real number and $d \geq 0$. For any $\epsilon > 0$, we follow the definition in [33]

$$H_\epsilon^d(S) = \inf\left\{\sum_{i=1}^{\infty} |U_i|^d : \{U_i\} \text{ is a } \epsilon - \text{cover of } S\right\}, \quad (1)$$

where the symbol inf $\{.\}$ indicates the infimum of $\{.\}$. As ϵ decreases, the class of permissible covers of S in Equation (1) is also reduced. Consequently, the infimum $H_\epsilon^d(S)$ increases, and so approaches a limit as $\epsilon \to 0$. Therefore, we can write

$$\begin{aligned} H^d(S) &= \lim_{\epsilon\to 0} H_\epsilon^d(S) \\ &= \lim_{\epsilon\to 0}\left[\inf\left\{\sum_{i=1}^{\infty} |U_i|^d : \{U_i\} \text{ is a } \epsilon - \text{cover of } S\right\}\right]. \end{aligned} \quad (2)$$

When $\epsilon \to 0$, the limit of $H_\epsilon^d(S)$ exists for any subset S of $\mathbb{R}^n$, and the limiting value can be 0 or ∞. We call $H^d(S)$ the *d-dimensional Hausdorff measure* of S [33].

H^d is a measure. Specifically, $H^d(\phi) = 0$, and if $E \subset S$ then $H^d(E) \leq H^d(S)$. If $\{S_i\}$ is any countable collection of Borel set, then

$$H^d\left(\bigcup_{i=1}^{\infty} S_i\right) = \sum_{i=1}^{\infty} H^d(S_i).$$

Hausdorff measures represent the familiar ideas of length, area and volume etc. in another form. The n-dimensional Hausdorff measure is just n-dimension Lebesgue measure, i.e. the n-dimensional volume, to within a constant multiple.

For any set S and $\epsilon < 1$, it is clear that $H_\epsilon^d(S)$ is a non-increasing function of d. According to Equation (2), it can be shown that $H^d(S)$ is also a non-increasing function of d. In fact, if $t > 0$ and $\{U_i\}$ is a ϵ-cover of S, we have

$$H^t_\epsilon(S) \le \sum_i |U_i|^t \le \epsilon^{t-d} \sum_i |U_i|^d. \tag{3}$$

Take the infimum, that is

$$H^t_\epsilon(S) \le \epsilon^{t-d} H^d_\epsilon(S).$$

Let $\epsilon \to 0$, if $H^d(S) < \infty$, then $H^t(S) = 0$ for $s < t$. Therefore, there exists a critical value of d, such that $H^d(S)$ jumps from ∞ to 0 at this point. This critical value is called the *Hausdorff Dimension* of S, and it is symbolized by $\dim_H S$.

Formally,

$$\begin{aligned} \dim_H S &= \inf\{d : H^d(S) = 0\} \\ &= \sup\{d : H^d(S) = \infty\}, \end{aligned} \tag{4}$$

and

$$H^d(S) = \begin{cases} \infty & \text{if } d < \dim_H S, \\ 0 & \text{if } d > \dim_H S. \end{cases}$$

If $d = \dim_H S$, probably $H^d(S)$ is 0 or ∞, or may satisfy

$$0 < H^d(S) < \infty.$$

2.1.2 Box Counting Dimension

Even though Hausdorff dimension is the oldest and probably the most important one, it has a major disadvantage that it is difficult to calculate or to estimate by computational methods in many cases. In practice, box counting dimension is convenient to apply. In this section, we will introduce box counting dimension [32]. It is one of the most widely used dimensions. It is relatively easy for mathematical calculation and empirical estimation, which induces large popularity.

At first, a measurement scale ϵ is adopted in the definition of dimension. For each ϵ, a set can be measured in a way that ignores irregularities of size less than ϵ, and a dimension is defined based on how these measurements change as $\epsilon \to 0$. Suppose S is a plane curve, the measurement $N_\epsilon(S)$ denotes the number of sets with length ϵ which divide the shape S [33]. A dimension of S is determined by the power law obeyed by $N_\epsilon(S)$ as $\epsilon \to 0$. If

$$N_\epsilon(S) \sim \mathcal{K}\epsilon^{-d}, \tag{5}$$

for constants $\mathcal{K}$ and d, we might say that S has dimension d, and $\mathcal{K}$ can be considered as "d-dimensional length" of S. Taking the logarithm of both sides in Equation (5) yields the formula:

$$\log_2 N_\epsilon(S) \simeq \log_2 \mathcal{K} - D \log_2 \epsilon,$$

in the sense that the difference of the two sides tends to 0 with ϵ, we have

$$d = \lim_{\epsilon \to 0} \frac{\log_2 N_\epsilon(S)}{\log_2(1/\epsilon)}. \tag{6}$$

Let S be a non-empty and bounded subset of $\mathbb{R}^n$, $\mho = \{\omega_i : i = 1, 2, 3, \ldots\}$ be covers of the set S. $N_\epsilon(S)$ denotes the number of covers, such that

$$N_\epsilon(S) = |\mho : D_i \le \epsilon|,$$

where D_i stands for the diameter of the ith cover. This equation means that $N_\epsilon(S)$ is the smallest number of subsets which cover the set S, and their diameters D_i's are not greater than ϵ. The upper and lower bounds of the box counting dimension of S can be defined by the following formulas:

$$\underline{\dim_B} S = \liminf_{\epsilon \to 0} \frac{\log_2 N_\epsilon(S)}{-\log_2 \epsilon}, \tag{7}$$

$$\overline{\dim_B} S = \overline{\lim_{\epsilon \to 0}} \frac{\log_2 N_\epsilon(S)}{-\log_2 \epsilon}, \tag{8}$$

where the over line stands for the upper bound of dimension while the under line for lower bound.

If both the upper bound $\overline{\dim_B} S$ and the lower bound $\underline{\dim_B} S$ are equal, i.e.

$$\liminf_{\epsilon \to 0} \frac{\log_2 N_\epsilon(S)}{-\log_2 \epsilon} = \overline{\lim_{\epsilon \to 0}} \frac{\log_2 N_\epsilon(S)}{-\log_2 \epsilon},$$

the common value is called *box counting dimension* or *box dimension* of S, namely:

$$\dim_B S = \lim_{\epsilon \to 0} \frac{\log_2 N_\epsilon(S)}{-\log_2 \epsilon}. \tag{9}$$

There exist five equivalent definitions of the box counting dimension, that can be found in Theorems 1 and 2 [34]:

Theorem 1. *Let S be a non-empty and bounded set in $\mathbb{R}^n$, and the upper box dimension $\overline{\dim_B} S$, lower box dimension $\underline{\dim_B} S$ and box dimension $\dim_B S$ be represented by:*

$$\begin{aligned} \underline{\dim_B} S &= \liminf_{\epsilon \to 0} \frac{\log_2 N_\epsilon(S)}{-\log_2 \epsilon}, \\ \overline{\dim_B} S &= \overline{\lim_{\epsilon \to 0}} \frac{\log_2 N_\epsilon(S)}{-\log_2 \epsilon}, \\ \dim_B S &= \lim_{\epsilon \to 0} \frac{\log_2 N_\epsilon(S)}{-\log_2 \epsilon}. \end{aligned}$$

In the above definition, $N_\epsilon(S)$ can be considered as one of the following cases:

1) the minimum number of closed balls of radius ϵ that cover S;
2) the minimum number of cubes with side ϵ that cover S;
3) the minimum number of sets with diameter D that cover S such that $d \le \epsilon$;
4) the number of ϵ-mesh cubs which intersect S;
5) the maximum number of disjoint balls of radius ϵ with centers in S.

Theorem 2. *Let S be a non-empty and bounded set in $\mathbb{R}^n$, and it satisfies that $1 < H^d(S)$ when $d = \dim_H S$, we have*

$$\dim_H S \le \underline{\dim_B} S \le \overline{\dim_B} S.$$

2.1.3 Minkowski Dimension

To facilitate the application of the box dimension to digital images, we will introduce Minkowski dimension [32] which is suitable for processing the digital images in computers.

Let S be a non-empty and bounded set in $\mathbb{R}^n$. For a constant d, if $\epsilon \to 0$, the limit of $Vol^n(S_\epsilon)/\epsilon^{n-d}$ is positive and bounded, we say that S has d dimension of *Minkowski dimension*, and is symbolized by $\dim_M S$. Here, $Vol^n(S_\epsilon)$ is called *Lebesgue Measure*. The relationship between the box dimension and Minkowski dimension can be provided by the following theorems, which can be found in [33].

Theorem 3. *Let S be a non-empty and bounded set in $\mathbb{R}^n$. Then we have*

$$\underline{\dim_B} S = n - \overline{\lim_{\epsilon \to 0}} \frac{\log_2 Vol^n(S_\epsilon)}{\log_2 \epsilon},$$

$$\overline{\dim_B} S = n - \liminf_{\epsilon \to 0} \frac{\log_2 Vol^n(S_\epsilon)}{\log_2 \epsilon},$$

where S_ϵ stands for ϵ-parallel body of S, and $Vol^n(S_\epsilon)$ denotes n-dimensional volume of S_ϵ.

From Theorem 3, we can derive Theorem 4, which is a very important theorem. It shows that the box counting dimension is equal to the Minkowski dimension in a non-empty and bounded set in $\mathbb{R}^n$.

Theorem 4. *Let S be a non-empty and bounded set in $\mathbb{R}^n$. We have*

$$\dim_B S = \dim_M S.$$

2.2 Approximation Algorithms of Fractal Dimension Calculation

In the applications of 1-dimensional signal processing and 2-dimensional image processing, we can only assign limited values to the scaling factor due to the discrete property. Therefore, the fractal dimension is usually calculated by some specific approximation algorithms. Even though the box counting dimension can be calculated, the procedure is highly time consuming [35]. Therefore several approximation algorithms such as Katz's, Petrosian's, and Higuchi's algorithms are invented. All of these algorithms have advantages and disadvantages and using them depends on the application [21]. Katz's algorithm [36] compared with Petrosian's algorithm [37] is slightly slower. Hiaguchi's [38], however, unlike the Petrosian's algorithm, it does not exist any pre-processing level for creating binary sequence. Hiaguchi's can be implemented directly on the analyzed signal.

2.2.1 Katz's Algorithm

Katz's FD calculation [36] eliminates the preprocessing step, and derived directly from the waveform. Katz's FD of a curve is defined as:

$$D = \frac{\log(L)}{\log(d)}, \tag{10}$$

where L is the sum of the distance of successive points, and d is the diameter estimated as the distance between the first point of the sequence and the farthest point. They are defined as:

$$L = \sum_{i=1}^{N-1} distance(i, i+1), \tag{11}$$

$$d = \max(distance(1, i)), \quad i = 1 \ldots N, \tag{12}$$

where N is the length of the signal. Then a normalization term is added to Equation (10) distance in Equation (10) by this average results in:

$$D = \frac{\log(\frac{L}{a})}{\log(\frac{d}{a})}, \tag{13}$$

where $a = L/n$. n is the number of steps of a signal (total values minus 1). Equation (10) can be written as:

$$D = \frac{\log(n)}{\log(\frac{d}{L}) + \log(n)}. \tag{14}$$

If n and L are fixed, a larger d indicates smaller fractal dimension because the total length of the signal is unchanged but the maximal span of it increases. That means the signal contains less detail parts so that its fractal dimension is smaller. On the other hand, if n and d are fixed, a larger L indicates larger fractal dimension because the maximal span is fixed, the only way to increase the total length of the signal is to add detail parts.

2.2.2 Petrosian's Algorithm

Petrosian applied a fast fractal dimension estimation based on Katz's FD [37]. A signal is recorded by subtracting continuous values from the original signal. Due to the sequence of subtractions, a binary sequence is created assigning +1 if the subtraction result is positive, or -1 if the result is negative, respectively. The FD is then computed as:

$$D = \frac{\log(n)}{\log(n) + \log\left(\frac{n}{n + 0.4N_\Delta}\right)}, \tag{15}$$

where n is the number of steps of a signal (total values minus 1) and N_Δ is the number of sign changes in the binary sequence.

The computational cost of Petrosian's FD is low since it only calculate the changes of successive values. It's obvious that in Equation (15), a larger N_Δ means a larger fractal dimension, because larger N_Δ means more complex curve.

2.2.3 Higuchi's Algorithm

Higuchi proposed a new approach to estimate the fractal dimension [38], which is the so-called Higuchi's fractal dimension (HFD). Assume an one-dimension signal $\{x_i\}, i = 1...N$. By defining a measurement scale k, the signal x can be separated as:

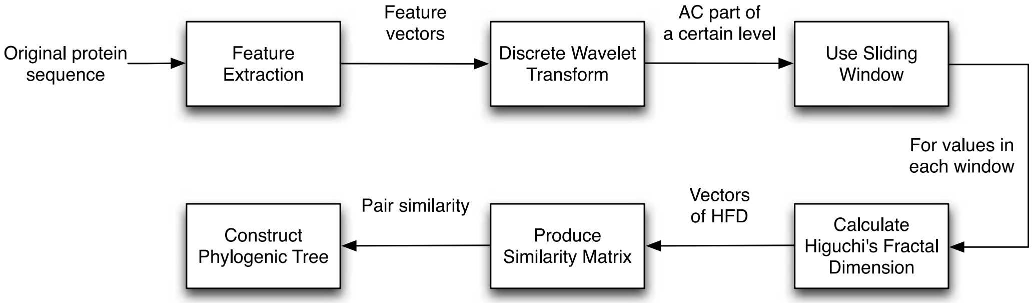

Original protein sequence
Feature Extraction
Feature vectors
Discrete Wavelet Transform
AC part of a certain level
Use Sliding Window
For values in each window
Calculate Higuchi's Fractal Dimension
Vectors of HFD
Produce Similarity Matrix
Pair similarity
Construct Phylogenic Tree

Fig. 1. Flowchart of our proposed algorithm HWF.

$$x_m^k = \left\{ x(m), x(m+k),\ x(m+2k), \ldots, \right. \\ \left. x\left(m + \left\lfloor \frac{N-m}{k} \right\rfloor k\right)\right\},\ m = 1 \ldots k, \tag{16}$$

where $\lfloor \cdot \rfloor$ is the floor operation, m is the starting position and $\lfloor \frac{N-m}{k} \rfloor$ indicates how many terms in x_m^k.

Then we can calculate the approximated length by measurement scale k and started from m:

$$L_m(k) = \left(\sum_{i=1}^{\lfloor \frac{N-m}{k} \rfloor} |x\big(m+ik\big) - x\big(m+(i-1)k\big)|(N-1) \right) \Bigg/ \\ \left(\left\lfloor \frac{N-m}{k} \right\rfloor k^2 \right),\ m = 1 \ldots k, \tag{17}$$

where $(N-1)/(\lfloor \frac{N-m}{k} \rfloor k)$ is the normalization term. Dividing by $\lfloor \frac{N-m}{k} \rfloor k$ is to calculate the average difference between each sample in the signal. After multiplying $N-1$ in the numerator, it is the approximated length of the signal by given measurement scale k.

At last, the average length of the signal is given by:

$$L(k) = \frac{1}{k} \sum_{m=1}^{k} L_m(k), \tag{18}$$

which is under the measurement scale k.

While given $k = 1...K$, where K is a pre-defined maximal value of k, the fractal dimension d^* of the signal which is also the slope of $\log(L(k))$ and $\log(\frac{1}{k})$ can be evaluated by a simple least square method:

$$d^* = \arg\min_d \sum_{k=1}^{K} \left(d \log\left(\frac{1}{k}\right) - \log\big(L(k)\big) + c \right)^2, \tag{19}$$

where c is the bias.

To sum up, the traditional box counting dimension can calculate precise fractal dimension. However, it needs the signal or image to be thresholded. Different threshold value will produce different fractal dimension. Therefore the most frequently used fractal dimension calculation methods are actually the Katz's, Petrosian's and Higuchi's algorithms as mentioned before. Katz's algorithm is a fast edition of Petrosian's. Both of them can approximate the fractal dimension without using the scaling value. Higuchi's algorithm can get the more accurate result than the other two algorithms, because it uses the scaling value as same as box counting algorithm. In this paper, our proposed protein comparison algorithm will adopt Higuchi's algorithm as the calculation of fractal dimension.

3 Proposed Approach

In this paper, the Hybrid method involving discrete Wavelet transform, Fractal dimension calculation (HWF) with sliding window are proposed to construct and analyze the feature vector of a protein sequence. After selection of a proper protein property, the protein sequences are described by using DWT. Then a technique called sliding window is used to deal with the values within a window with fixed length of a signal. For values in each fixed window, the Higuchi's fractal dimension is calculated. Based on the similarity matrix of the feature vector, a set of protein sequences are clustered into different groups to build a phylogenic tree. An overall description of the proposed approach can be illustrated in Fig. 1.

3.1 Feature Extraction

Even though most methods for exploring the sequence collections in databases are based on sequence comparison, other methods based on the analysis of physicochemical properties of the amino acids have also been proven to be useful [4]. Among the various properties of amino acids, hydropathy is known to be well conserved, this can be explained by the fact that the hydrophobic residues contribute significantly to the folding of globular domains and other structural elements of proteins. It has been generally understood that the hydrophobic effect of amino acid residues is the driving force responsible for the folding of protein. Hydropathy distribution has been used to detect analogous as well as distantly related protein [39]. The hydropathy distribution for the protein sequence has been recognized as a primary feature for characterization of protein structure in the term of hydropathy profiles. [40] illustrates that the hydropathy analysis using such profiles was used in classification of new protein sequence data. Hydropathy analysis has also been used for estimation of membrane protein structure [42] and to analyze protein sequence databases [43].

There are 20 kinds of amino acids joined by peptide bonds to form the protein. In general, they are presented as capital letters: A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W and Y. The proposed approach is based on the extraction of specific hydropathy features from the protein sequence. The one-to-one mapping of amino acids and their

TABLE 1
Feature Values of Each Amino Acid

Name	Abbreviation	code	H-value
Alanine	Ala	A	1.8
Cysteine	Cys	C	2.5
Aspartic acid	Asp	D	–3.5
Glutamic acid	Glu	E	–3.5
Phenylalanine	Phe	F	2.8
Glycine	Gly	G	–0.4
Histidine	His	H	–3.2
Isoleucine	lle	I	4.5
Lysine	Lys	K	–3.9
Leucine	Leu	L	3.8
Methionine	Met	M	1.9
Aspargine	Asp	N	–3.5
Proline	Pro	P	–1.6
Glutamine	Glu	Q	–3.5
Arginine	Arg	R	–4.5
Serine	Ser	S	–0.8
Threonine	Thr	T	–0.7
Valine	Val	V	4.2
Tryptophan	Trp	W	–0.9
Tyrosine	Tyr	Y	–1.3

hydropathy values are shown in Table 1. A vectorial representation of the protein sequence is achieved by using the features.

3.2 Discrete Wavelet Transform

Discrete wavelet transform is a relative new mathematical tool which is similar to discrete Fourier transform. However it can analyze a signal in both time and frequency. It is an efficient tool to decompose a signal into two parts: Approximate Coefficients (AC) and Detail Coefficients (DC).

Let $\phi(x)$ is a scaling function:

$$\phi(x) = \sqrt{2}\sum_{n\in Z} h_n \phi(2x-n), \tag{20}$$

where Z is a set of integers and $\{h_n\}$ denote low-pass filters. By using the scaling function $\phi(x)$, then we can construct the wavelet function $\psi(x)$:

$$\psi(x) = \sqrt{2}\sum_{n\in Z} g_n \phi(2x-n), \tag{21}$$

where $\{g_n\}$ denote high-pass filters. In the wavelet transform, we can assume that the shifted scaling functions $\{\phi(x-k)\}$ and the shifted wavelet functions $\{\psi(x-k)\}$ are both orthonormal.

Let $\{a_k^0\}$ be the original signal, we can decompose the signal into AC and DC in the next level:

$$a_k^1 = \sum_{i\in Z} a_i^0 \overline{h}_{i-2k}, \tag{22}$$

and

$$d_k^1 = \sum_{i\in Z} a_i^0 \overline{g}_{i-2k}. \tag{23}$$

The decomposition can be operated level by level:

$$a_k^{j+1} = \sum_{i\in Z} a_i^j \overline{h}_{i-2k}, \quad j = 0, 1, 2, \dots, \tag{24}$$

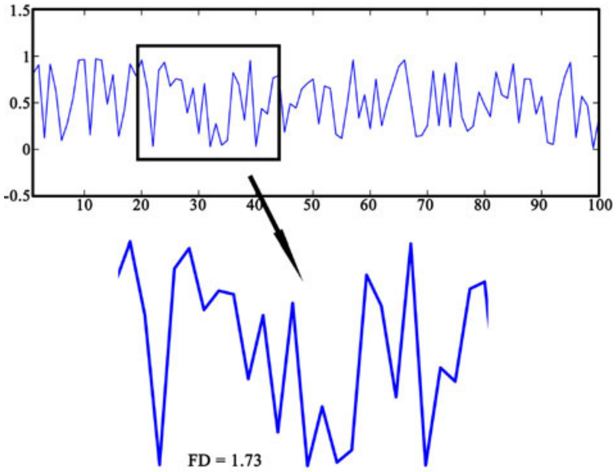


Fig. 2. An example of sliding window.

and

$$d_k^{j+1} = \sum_{i\in Z} a_i^j \overline{g}_{i-2k}, \quad j = 0, 1, 2, \dots \tag{25}$$

Note that the length of the coefficients after decomposition, namely AC and DC, is always only a half of the length of the previous level sequence. By using the DWT to process protein sequences, AC can describe the overall trend and general characteristics of signals, while DC describes some local details. In the proposed method, the AC of Haar wavelet is used, because it is simple and effective. It has been extensively used for protein sequence analysis [13], [44].

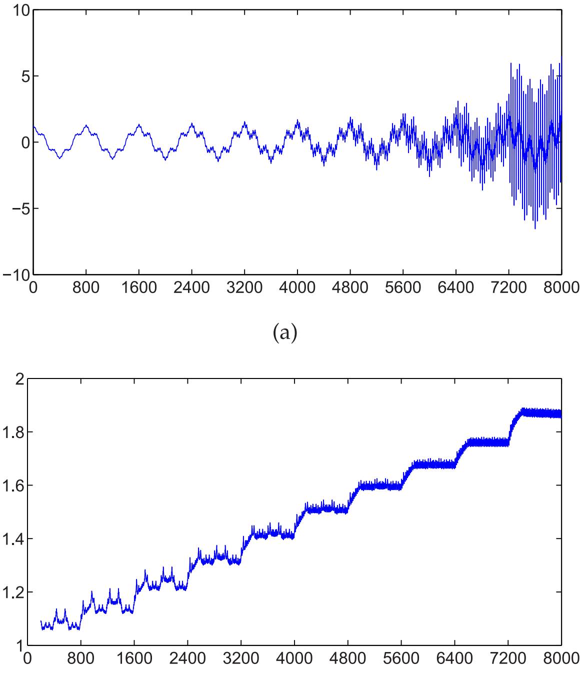
3.3 Windowed Fractal Dimension

If the Higuchi's fractal dimension is directly applied to calculate the fractal dimension of the wavelet coefficients, only a scalar will be produced. It is insufficient to represent the protein sequence with enough discriminance. Sliding window [45] is a technique which only deals with the values within the fixed length window of a signal. In our proposed algorithm HWF, we calculate the Higuchi's fractal dimension of values within a window which moves along the signal. Fig. 2 shows the principle of sliding window. Combined with Higuchi's fractal dimension, sliding window can produce a feature vector with length $N-w+1$ other than only a scalar value, where N is the length of the original signal and w is the window width. We can derive Equation (17) to

$$L_m^j(k) = \sum_{i=1}^{\lfloor \frac{w-m}{k} \rfloor} |x(m+ik+j-1) - x(m+(i-1)k+j-1)|(w-1) \Big/ \left\lfloor \frac{w-m}{k} \right\rfloor k^2, \tag{26}$$

where $m = 1 \dots k,\ j = 1 \dots N-w+1$. Then calculate $L(k)$ for each window j:

$$L^j(k) = \frac{1}{k}\sum_{m=1}^{k} L_m^j(k). \tag{27}$$



(a)

(b)

Fig. 3. (a) A simulated signal connecting 10 Weierstrass cosine functions with increasing H value for every 1,000 samples, (b) Higuchi's fractal dimension of (a) using a sliding window with width 200.

At last, the Higuchi's fractal dimension of each window are

$$d^{j*} = \arg\min_{d} \sum_{k=1}^{K} \left(d \log\left(\frac{1}{k}\right) - \log\left(L^j(k)\right) + c \right)^2. \tag{28}$$

A vector of Higuchi's fractal dimension $D^* = \{d^{1*}, d^{2*}, d^{3*}, \ldots, d^{(N-w+1)*}\}$.

We use a synthetic signal call Weierstrass cosine function as example to illustrate Higuchi's fractal dimension and sliding window. Weierstrass cosine function [41] is as follows:

$$W_H(t) = \sum_{i=0}^{M} \lambda^{-iH} \cos\left(2\pi\lambda^i t\right), \tag{29}$$

where $H \in (0,1)$ and the theoretical fractal dimension of the Weierstrass cosine function is $2 - H$. We set $\lambda = 5$ and $M = 26$ following the construction in [28]. $t \in (0,1)$ is separated into 1,000 samples with step 0.001. We construct 10 curves with different H from 1 decreasing to 0 with step 0.1, and the theoretical fractal dimensions of these 10 curves are from 1 to 2 with step 0.1. A window with width 200 is adopted in this example. Fig. 3 shows the plots of 10 Weierstrass cosine function with different H values and their corresponding HFD.

3.4 Similarity Measure

The key procedure in protein sequence comparison is to choose a proper similarity function. There are some similarity measures which are commonly used for protein comparison. e.g. the cosine similarity [54], Jaccard index [53], The Tanimoto coefficient [54] and squared Euclidean distance. In this paper, the cosine similarity is adopted in the proposed method:

$$C(x,y) = \frac{< x, y >}{||x||||y||}, \tag{30}$$

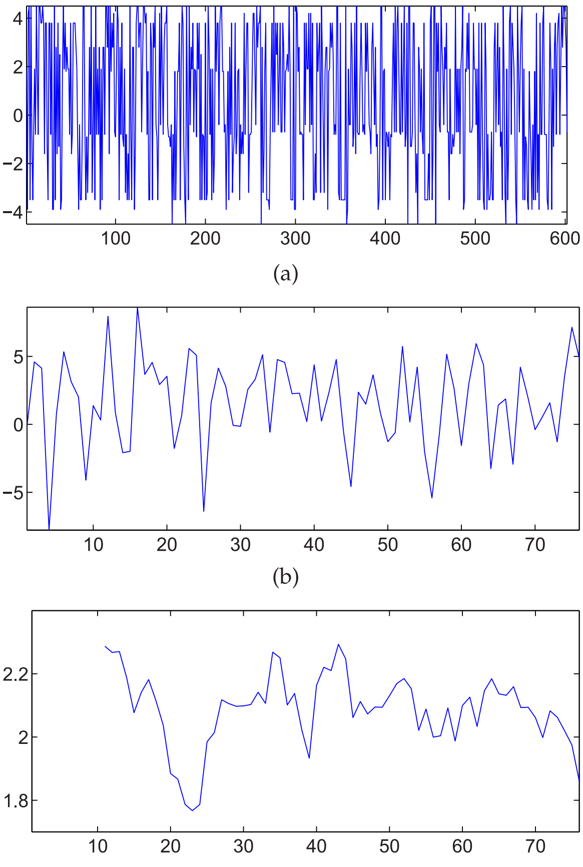
where $< \cdot, \cdot >$ is the inner product. It calculates the angle between two normalized vectors. Cosine similarity has a special property that the resulting similarity measure always falls into the range of -1 and $+1$, where $+1$ means that the two vectors are exactly matching, 0 means they are orthogonal and -1 means they matches but in opposite direction.

TABLE 2
Nine ND5 Protein Sequences from NCBI Database

Source	Accession ID	Length
Human	AP-000649	603
Common chimpanzee	NP-008196	603
Pigmy chimpanzee	NP-008209	603
Gorilla	NP-008222	603
Fin whale	NP-006899	606
Blue whale	NP-007066	606
Mouse	NP-904338	607
Rat	AP-004902	610
Opossum	NP-007105	602

3.5 Algorithm Summary

Algorithm 1 summarizes the calculation of the similarity between protein sequences based on DWT decomposition and HFD with sliding window.



(a)

(b)

(c)

Fig. 4. (a) Hydropathy value of the human protein sequence (Accession ID: AP-000649 in NCBI), (b) the Third level discrete wavelet transform, (c) the Higuchi's fractal dimension of (b) using sliding window with width 11.

TABLE 3
The Distance Matrix of Nine ND5 Protein Sequences Generated by HWF

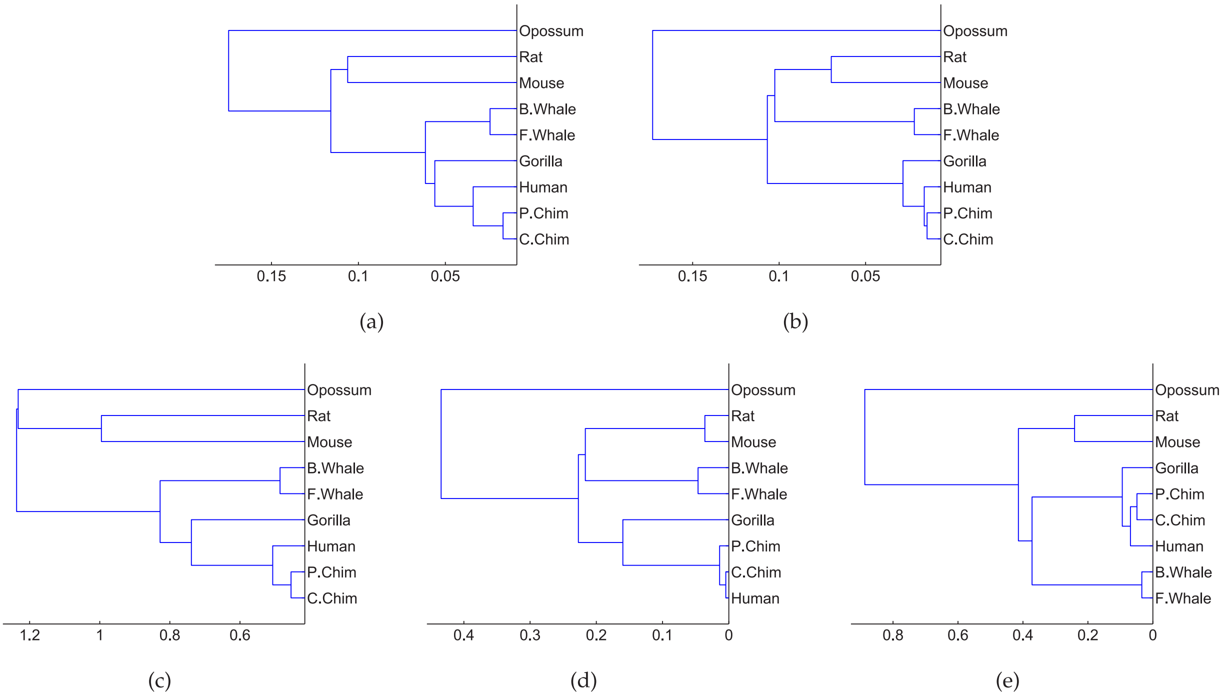
	Human	Gorilla	C.Chim	P.Chim	F.Whale	B.Whale	Mouse	Rat	Opossum
Human	0	0.1062	0.0343	0.0340	0.1348	0.1820	0.1593	0.1367	0.1747
Gorilla		0	0.0768	0.0561	0.0628	0.0615	0.1159	0.2170	0.2541
C.Chim			0	0.0168	0.1467	0.1491	0.1327	0.1673	0.1944
P.Chim				0	0.1057	0.1240	0.1268	0.1620	0.1798
F.Whale					0	0.0243	0.1404	0.1600	0.1888
B.Whale						0	0.1302	0.1902	0.2237
Mouse							0	0.1062	0.2355
Rat								0	0.2483
Opossum									0

4 Experiments and Results

In this section, nine ND5 (NADH dehydrogenase subunit 5) protein sequences are selected from the NCBI protein cluster data sets in the experiment. They have nearly same length since they are homologous proteins in organelle genome group. The information of the used protein sequences is shown in Table 2. ND5 is chosen because the protein sequences in this data set have similar length that our proposed method can be directly used. Fig. 4 shows the DWT and HFD on a human protein sequence feature using our proposed HWF. The width of the sliding window is set to 11 and the wavelet level M is set to 3 using trial and error method by observing the generated phylogenetic tree.

The distance matrix of nine ND5 protein sequences generated by our proposed HWF is shown in Table 3. We also compare our proposed algorithm with Su's model [46], Zhang's model [47], Yao's model [48] and MEGA software [49], as shown in Fig. 5. The phylogenetic tree generated by our propped HWF is created by dendrogram function and linkage function in the statistics toolbox in Matlab by given the distance matrix. In MEGA software, we generate the phylogenetic tree by maximum likelihood method.

In Fig. 5, the abscissa is the similarity between each species and the ordinate is the 9 ND5 species. The proposed HWF shown in Fig. 5a is consistent with the result generated from MEGA software shown in Fig. 5e and some known facts of evolution [50], [51], [52]. Su's model shown in Fig. 5b connects the cluster of rat and mouse to the cluster of blue whale and fin whale. However, the evolution fact shows that the cluster of whales should connects to the cluster of the primate. Zhang's model shown in Fig. 5c connects the cluster of rat and mouse to opossum first. However opossum is the known farthest species to other eight kinds. Yao's model shown in Fig. 5d clusters Human and common chimpanzee first, instead of the two kinds of chimpanzee. Therefore it's the worst result within the above models. Comparing with HWF, these methods use simple feature transform techniques and then apply a distance metric to get the distance



(a) (b) (c) (d) (e)

Fig. 5. The phylogenetic tree of nine protein sequences constructed by (a) HWF, (b) Su's model, (c) Zhang's model, (d) Yao's model, (e) MEGA software.

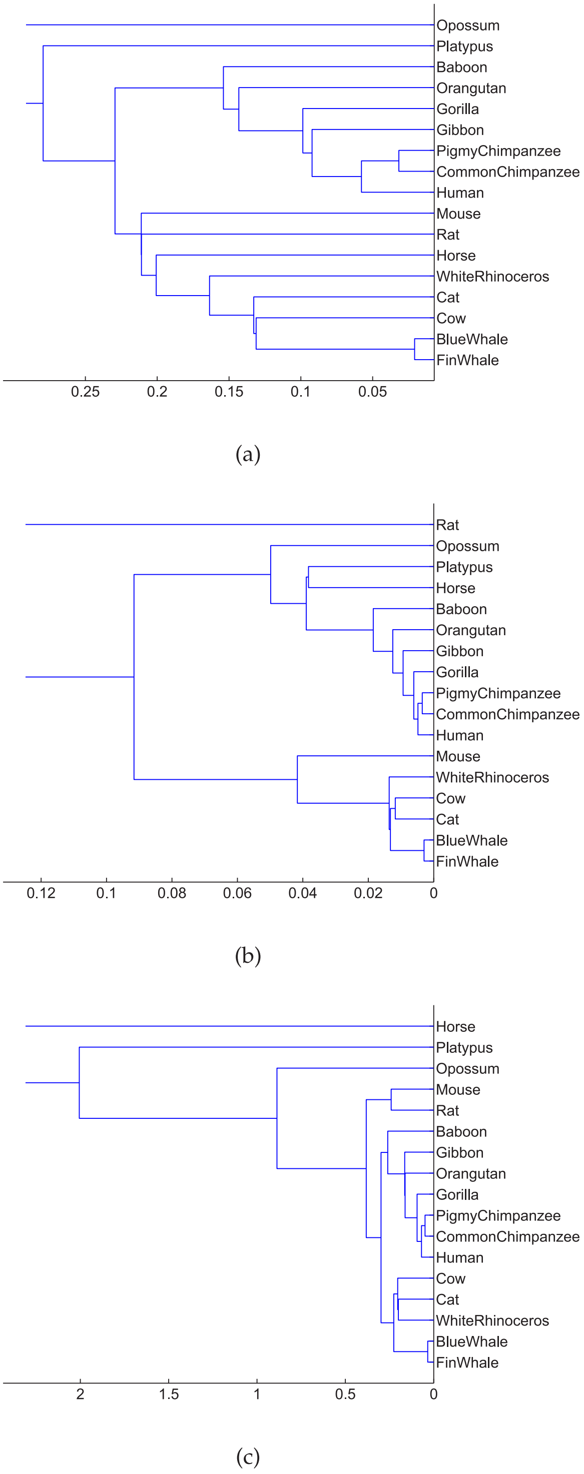


Fig. 6. The phylogenetic tree of 17 ND5 protein sequences constructed by (a) HWF, (b) Su's model, (c) MEGA software.

matrix. Therefore, they are less accurate to represent the potential information of protein sequences so that the resulting phylogenetic trees are divergent from the known fact. HWF uses a more sophisticated way which incorporates the fractal dimension of signal to analyze protein sequences. Even though the same distance metric is adopted as other methods, the features extracted by HWF can better reflect the essence of the sequence information.

TABLE 4
Seventeen ND5 Protein Sequences from NCBI Database

Source	Accession ID	Length
Human	AP-000649	603
Common chimpanzee	NP-008196	603
Pigmy chimpanzee	NP-008209	603
Gorilla	NP-008222	603
Orangutan	NP-008235	603
Gibbon	NP-007832	603
Baboon	NP-008468	603
Horse	NP-007170	604
White rhinoceros	NP-007443	606
Cat	NP-008261	606
Fin whale	NP-006899	606
Blue whale	NP-007066	606
Cow	YP-209215	606
Mouse	NP-904338	607
Rat	AP-004902	610
Opossum	NP-007105	602
Platypus	NP-008053	604

Algorithm 1. Hybrid of Discrete Wavelet Transform and Higuchi's Fractal Dimension for Protein Sequences Comparison

```
Input: N protein sequences $\{P_i,\ i = 1 \dots N\}$.
Output: Distance matrix $T$.
  1) Map each amino acid in $P_i$ to its corresponding hydropathy value shown in Table 1 to generate the feature vectors $\{S_i,\ i = 1 \dots N\}$ with same length by cutting the tails.
  2) Calculate the DWT by using Haar wavelet function of $\{S_i,\ i = 1 \dots N\}$ at level M:
        $(AC_i, DC_i) = DWT(S_i, Haar, M), i = 1 \dots N.$   (31)
  3) Given window width $w$, slide the window along $\{AC_i,\ i = 1 ... N\}$ to generate $K_i$ sequences for each $AC_i$:
        $W_i^{(k)} = [AC_i(k), AC_i(k+1), \dots, AC_i(k - w + 1)],\ k = 1 \dots K_i,\ i = 1 \dots N,$   (32)
     where $K_i = \text{length}\,(AC_i) - w + 1$.
  4) Calculate the Higuchi's fractal dimension for values in every window:
        $H_i(k) = HFD\big(W_i^{(k)}\big),\ k = 1 \dots K_i,\ i = 1 \dots N,$   (33)
     and
        $H_i = [H_i(1), H_i(2), \dots, H_i(K_i)],\ i = 1 \dots N.$   (34)
  5) Calculate the pair distance of $\{H_i,\ i = 1 \dots N\}$:
        $d(H_i, H_j) = 1 - \frac{< H_i, H_j >}{|H_i| \cdot |H_j|}, i = 1 \dots N,\ j = 1 \dots N.$   (35)
  6) Construct the distance matrix $T$ where the element in the $i$th row and the $j$th column is $d(H_i, H_j)$.
```

Besides, another experiment which contains more protein sequences is shown in Fig. 6, where 17 protein sequences are compared in this experiment. The

TABLE 5
The Distance Matrix of 17 ND5 Protein Sequences Generated by HWF

	Human	C.Chim	P.Chim	Gorilla	Oran.	Gibbon	Baboon	Horse	W.Rhino.	Cat	F.Whale	B.Whale	Cow	Mouse	Rat	Opossum	Platypus
Human	0	0.0578	0.0754	0.1334	0.1559	0.0922	0.2174	0.3748	0.3013	0.2765	0.3561	0.3279	0.3228	0.3601	0.3373	0.3044	0.4352
C.Chim	0	0	0.0318	0.1147	0.1432	0.0994	0.1539	0.3385	0.2952	0.2900	0.3695	0.3329	0.3066	0.3325	0.2941	0.3190	0.4480
P.Chim	0	0	0	0.0986	0.1629	0.1179	0.1709	0.3793	0.3062	0.3072	0.4214	0.3751	0.3412	0.3230	0.3089	0.2937	0.4384
Gorilla	0	0	0	0	0.2419	0.1467	0.2372	0.4211	0.3935	0.2823	0.4434	0.4024	0.3780	0.4004	0.3620	0.3742	0.5748
Oran.	0	0	0	0	0	0.1900	0.1874	0.3584	0.2365	0.2294	0.3616	0.3326	0.3206	0.3659	0.2587	0.2986	0.3031
Gibbon	0	0	0	0	0	0	0.2650	0.3133	0.3127	0.2361	0.2943	0.2817	0.2328	0.3337	0.3194	0.3001	0.4602
Baboon	0	0	0	0	0	0	0	0.3583	0.3332	0.3247	0.4327	0.3654	0.3712	0.3622	0.2961	0.2944	0.3595
Horse	0	0	0	0	0	0	0	0	0.2548	0.2529	0.2227	0.2220	0.2007	0.3699	0.2658	0.3259	0.4193
W.Rhino.	0	0	0	0	0	0	0	0	0	0.1636	0.1862	0.1728	0.2201	0.3207	0.3298	0.3105	0.3863
Cat	0	0	0	0	0	0	0	0	0	0	0.1513	0.1495	0.1328	0.2743	0.2111	0.3012	0.3467
F.Whale	0	0	0	0	0	0	0	0	0	0	0	0.0208	0.1311	0.2974	0.3718	0.4470	0.5173
B.Whale	0	0	0	0	0	0	0	0	0	0	0	0	0.1339	0.2506	0.3445	0.4161	0.4694
Cow	0	0	0	0	0	0	0	0	0	0	0	0	0	0.2110	0.2517	0.3555	0.4753
Mouse	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.3286	0.3672	0.5332
Rat	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.4053	0.2794
Opossum	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.3593
Platypus	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

information of the used protein sequences is shown in Table 4. In this experiment, our proposed model HWF is much closer to the known fact [52]. The order from inside to outside in [52] is gorilla, orangutan and gibbon, but the order in our method is gibbon, gorilla and orangutan. Besides, cat and horse should be connected first and then to the group of whales and cow. And the group of mouse the rat should be one more level out. However, comparing to the result generated by Su's model in Fig. 6b, our method is much better. In Su's model, the platypus and horse are classified into a group and connects to the primates. Mouse is separated from rat and classified into the group of other mammals. Rat becomes the farthest species against other instead of opossum. In addition, we also compare to the result generated by MEGA software using maximum likelihood shown in Fig. 6c. It is also different from the fact of evolution. The horse is classified as the farthest species against the others. The distance matrix of 17 ND5 protein sequences generated by our proposed HWF is shown in Table 5.

5 Conclusion

In order to analyze a large amount of protein sequence data precisely and effectively, this paper introduces a hybrid model which is based on wavelet transform and fractal dimension, to measure the similarity of protein sequence using hydropathy characteristic.

The presented features are based on hydropathy properties of amino acids sequences. The feature was designed to formalize significant parts of protein sequences structurally and functionally. The main contribution is that we proposed a hybrid model, e.g. the HWF model, which is based on wavelet transform and fractal dimension to create a new representation of protein sequence, and measure the sequence similarity. The fractal dimension can capture important characteristic of fractals that contains information about their geometrical structure. The proposed approach is based on such characteristics. The experimental results illustrate that the performance of HWF model is better than that of Su' model, Zhang's model, Yao's model and MEGA software, and is consistent with the known fact.

Acknowledgments

This work was supported by the Research Grants MYRG205 (Y1-L4)-FST11-TYY, MYRG187(Y1 -L3)- FST11- TYY and Chair Prof. Grants RDG009/FST-TYY of University of Macau as well as Macau FDC grants T-100-2012-A3 and 026-2013-A. This research project was also supported by the National Natural Science Foundation of China 61273244.

References

[1] F. Shi, Q. Chen, and X. Niu, "Functional similarity analyzing of protein sequences with empirical mode decomposition," in *Proc. 4th Int. Conf. Fuzzy Syst. Knowl. Discov.*, vol. 2, pp. 766–770, 2007.

[2] Y. M. Cheung and H. Jia, "Categorical-and-numerical-attribute data clustering based on a unified similarity metric without knowing cluster number," *Pattern Recog.*, vol. 46, no. 8, pp. 2228–2238, 2013.

[3] K. Tsuda, H. Shin, and B. Scholkopf, "Fast protein classification with multiple networks," *Bioinformatics*, vol. 21, no. 2, pp. 59–65, 2005.

[4] D. S. Huang, X.-M. Zhao, G.-B. Huang, and Y.-M. Cheung, "Classifying protein sequences using hydropathy blocks," *Pattern Recog.*, vol. 39, no. 12, pp. 2293–2300, 2006.

[5] D. J. Rigden, *From Protein Structure to Function with Bioinformatics*. New York, NY: Springer-Verlag, 2009.

[6] C. H. Trad, Q. Fang, and I. Cosic, "Protein sequence comparison based on the wavelet transform approach," *Protein Eng.*, vol. 15, no. 3, pp. 193–203, 2002.

[7] Y. Shibberu and A. Holder, "A spectral approach to protein structure alignment," *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, vol. 8, no. 4, pp. 867–875, Jul. 2011.

[8] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, "Gapped blast and psi-blast: A new generation of protein database search programs," *Nucleic Acids Res.*, vol. 25, no. 17, pp. 3389–3402, 1997.

[9] A. Krogh, M. Brown, I. S. Mian, K. Sjölander, and D. Haussler, "Hidden Markov models in computational biology: Applications to protein modeling," *J. Mol. Biol.*, vol. 235, pp. 1501–1531, 1994.

[10] M. Bhasin and G. P. S. Raghava, "GPCRpred: An SVM-based method for prediction of families and subfamilies of g-protein coupled receptors," *Nucleic Acids Res.*, vol. 32, pp. 383–389, 2004.

[11] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions and reversals," *Soviet Phys. Doklady*, vol. 10, pp. 707–710, Feb. 1966.

[12] A. Kelil, S. Wang, R. Brzezinski, and A. Fleury, "CLUSS: Clustering of protein sequences based on a new similarity measure," *BMC Bioinformatics*, vol. 8, article 286, 2007.

[13] P. Lio, "Wavelets in bioinformatics and computational biology: State of art and perspectives," *Bioinformatics*, vol. 19, no. 1, pp. 2–9, 2003.

[14] T. Meng, A. T. Soliman, M.-L. Shyu, Y. Yang, S.-C. Chen, S. Iyengar, J. Yordy, and P. Iyengar, "Wavelet analysis in current cancer genome research: A survey," *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, vol. 10, no. 6, pp. 1442–14359, Nov.–Dec. 2013.

[15] H. Hirakawa, S. Muta, and S. Kuhara, "The hydrophobic cores of proteins predicted by wavelet analysis," *Bioinformatics*, vol. 15, no. 2, pp. 141–148, 1999.

[16] J. Qiu, R. Liang, X. Zou, and J. Mo, "Prediction of protein secondary structure based on continuous wavelet transform," *Talanta*, vol. 61, no. 3, pp. 285–293, 2003.

[17] C. H. de Trad, Q. Fang, and I. Cosic, "Protein sequence comparison based on the wavelet transform approach," *Protein Eng.*, vol. 15, no. 3, pp. 193–203, 2002.

[18] Z.-N. Wen, K.-L. Wang, M.-L. Li, F.-S. Nie, and Y. Yang, "Analyzing functional similarity of protein sequences with discrete wavelet transform," *Comput. Biol. Chem.*, vol. 29, no. 3, pp. 220–228, 2005.

[19] H. Chen, F. Gu, and Z. Huang, "A compound method of protein secondary structure prediction and its implementation," in *Proc. First IEEE Int. Multi-Symp. Comput. Comput. Sci.*, vol. 1, 2006, pp. 104–109.

[20] C. Arizmendi, A. Vellido, and E. Romero, "Classification of human brain tumours from MRS data using discrete wavelet transform and Bayesian neural networks," *Expert Syst. Appl.*, vol. 39, no. 5, pp. 5223–5232, 2012.

[21] R. Esteller, G. Vachtsevanos, J. Echauz, and B. Lilt, "A comparison of fractal dimension algorithms using synthetic and experimental data," in *Proc. Int. Symp. Circuits Syst.*, vol. 3, 1999, pp. 199–202.

[22] T. Nelson and A. Goldberger, "Fractal electrodynamics of ventricular depolarization-effect of conduction defects," *Circulation*, vol. 82, pp. 739–739, 1990.

[23] L. M. Boxt, J. Katz, L. S. Liebovitch, R. Jones, P. D. Esser, and L. Reid, "Fractal analysis of pulmonary arteries: The fractal dimension is lower in pulmonary hypertension," *J. Thoracic Imaging*, vol. 9, no. 1, pp. 8–13, 1994.

[24] Q. Chen, Y. He, and B. Kang, "Fractal analyze protein structural similarity based on structural spectrum," in *Proc. WRI Global Congress Intell. Syst.*, 2009, pp. 165–169.

[25] Y. Tao, T. R. Ioerger, and J. C. Sacchettini, "Extracting fractal features for analyzing protein structure," in *Proc. 16th IEEE Int. Conf. Pattern Recog.*, vol. 2, 2002, pp. 482–485.

[26] T. Holden, N. Gadura, E. Cheung, P. Schneider, G. Tremberger, N. Elham, D. Sunil, D. Lieberman, and T. Cheung, "Nipah virus classification via fractal dimension & shannon entropy," in *Proc. IEEE 4th Int. Conf. Bioinformatics Biomed. Eng.*, 2010, pp. 1–4.

[27] H. Zhang and W. Kinsner, "Feature extraction from DNA sequences by multifractal analysis," in *Proc. Int. Conf. Eng. Med. Biol. Soc.*, vol. 2, 2001, pp. 1567–1572.

[28] R. Esteller, G. Vachtsevanos, J. Echauz, and B. Litt, "A comparison of waveform fractal dimension algorithms," *IEEE Trans. Circuits Syst. I: Fundam. Theory Appl.*, vol. 48, no. 2, pp. 177–183, Aug. 2001.

[29] Y. Tao, E. C. M. Lam, and Y. Y. Tang, "Feature extraction using wavelet and fractal," *Pattern Recog. Lett.*, vol. 22, no. 3, pp. 271–287, 2001.

[30] S. Gunasekaran and K. Revathy, "Fractal dimension analysis of audio signals for indian musical instrument recognition," in *Proc. Int. Conf. Audio, Lang. Image Process.*, 2008, pp. 257–261.

[31] H. Felix, "Dimension und äuβ eres Maβ," *Math. Ann.* vol. 79, pp. 157–179, 1918.

[32] S. M. Robert, *Fractals, Chaos, Power Laws: Minutes from an Infinite Paradise*. New York: NY, Dover, 2012.

[33] K. Falconer, *Fractal Geometry: Mathematical Foundations and Applications*. Hoboken: NJ, Wiley, 2007.

[34] Y. Y. Tang, H. Ma, D. Xi, X. Mao, and C. Y. Suen, "Modified fractal signature (MFS): A new approach to document analysis for automatic knowledge acquisition," *IEEE Trans. Knowl. Data Eng.*, vol. 9, no. 5, pp. 747–762, Sep. 1997.

[35] B. S. Raghavendra and N. D. Dutt, "Computing fractal dimension of signals using multiresolution box-counting method," *Int. J. Inform. Math. Sci.*, vol. 6, no. 1, pp. 50–65, 2010.

[36] M. J. Katz, "Fractals and the analysis of waveforms," *Comput. Biol. Med.*, vol. 18, no. 3, pp. 145–156, 1988.

[37] A. Petrosian, "Kolmogorov complexity of finite sequences and recognition of different preictal EEG patterns," in *Proc. 8th IEEE Symp. Comput.-Based Med. Syst.*, 1995, pp. 212–217.

[38] H. Tomoyuki, "Approach to an irregular time series on the basis of the fractal theory," *Phys. D: Nonlinear Phenomena*, vol. 31, no. 2, pp. 277–283, 1988.

[39] S. Y. Chung and S. Subbiah, "A structural explanation for the twilight zone of protein sequence homology," *Structure*, vol. 4, no. 10, pp. 1123–1127, 1996.

[40] X. Yu and D. H. Walker, "Sequence and characterization of an ehrlichia chaffeensis gene encoding 314 amino acids highly homologous to the NAD A Enzyme," *FEMS Microbiol. Lett.*, vol. 154, no. 1, pp. 53–58, 1997.

[41] C. Tricot, *Curves and Fractal Dimension*. New York: NY, Springer-Verlag, 1995.

[42] D. Slotboom and J. S. Lolkema, "Estimation of structural similarity of membrane proteins by hydropathy profile alignment," *Molecular Membrane Biol.*, vol. 15, no. 1, pp. 33–42, 1998.

[43] J. D. Clements and R. E. Martin, "Identification of novel membrane proteins by searching for patterns in hydropathy profiles," *Eur. J. Biochem.*, vol. 269, no. 8, pp. 2101–2107, 2002.

[44] M. Hayat and A. Khan, "Membrane protein prediction using wavelet decomposition and pseudo amino acid based feature extraction," in *Proc. Int. Conf. Emerging Technol.*, 2010, pp. 1–6.

[45] X. Gong and T. Corpetti, "Adaptive window size estimation in unsupervised change detection," *IEEE J. Select. Topics Appl. Earth Observ. Remote Sens.*, vol. 6, no. 2, pp. 991–1003, Nov. 2013.

[46] J. Su and J. Bao, "A wavelet transform based protein sequence similarity model," *Appl. Math*, vol. 7, no. 3, pp. 1103–1110, 2013.

[47] Y. Zhang and X. Yu, "Analysis of protein sequence similarity," in *Proc. 5th Int. Conf. Bio-Inspired Comput.: Theories Appl.*, 2010, pp. 1255–1258.

[48] Y. Yao, Q. Dai, C. Li, P. He, X. Nan, and Y. Zhang, "Analysis of similarity/dissimilarity of protein sequences," *Proteins: Struct., Function, and Bioinformatics*, vol. 73, no. 4, pp. 864–871, 2008.

[49] K. Tamura, J. Dudley, M. Nei, and S. Kumar, "MEGA4: Molecular evolutionary genetics analysis (MEGA) software version 4.0," *Molecular Biol. Evol.*, vol. 24, no. 8, pp. 1596–1599, 2007.

[50] M. Li, J. H. Badger, X. Chen, S. Kwong, P. Kearney, and H. Zhang, "An information-based sequence distance and its application to whole mitochondrial genome phylogeny," *Bioinformatics*, vol. 17, no. 2, pp. 149–154, 2001.

[51] V. Makarenkov and F. Lapointe, "A weighted least-squares approach for inferring phylogenies from incomplete distance matrices," *Bioinformatics*, vol. 20, no. 13, pp. 2113–2121, 2004.

[52] H. H. Otu and K. Sayood, "A new sequence distance measure for phylogenetic tree construction," *Bioinformatics*, vol. 19, no. 16, pp. 2122–2130, 2003.

[53] P. Jaccard, "Etude comparative de la distribution orale dans une portion des Alpes et des Jura," *Bull. del la Societe Vaudoise des Sci. Naturelles*, vol. 37, pp. 547–579, 1901.

[54] A. Tversky, "Features of Similarity," *Psychological Rev.*, vol. 84, no. 4, pp. 327–352, 1977.

Lina Yang (S'13) received the BS degree in computer engineering from Shijiazhuang Railway University, Hebei, China, and the MEng degree in computer science from the University of Malaya, Kuala Lumpur, Malaysia, in 2005 and 2011, respectively. She is currently working towards the PhD degree in the Faculty of Science and Technology, University of Macau, Macau, China. Her research interests include machine learning and image processing. She is a student member of IEEE.

Yuan Yan Tang (F'04) is a chair professor in the Faculty of Science and Technology at the University of Macau and a professor/adjunct professor/honorary professor at several institutes including Chongqing University in China, Concordia University in Canada, and Hong Kong Baptist University in Hong Kong. His current interests include wavelets, pattern recognition, and image processing. He has published more than 400 academic papers and is the author/coauthor of more than 25 monographs/books/bookchapters. He is the founder and editor-in-chief of the *International Journal on Wavelets, Multiresolution, and Information Processing* (*IJWMIP*), and associate editor of several international journals. He is the founder and chair of pattern recognition committee in IEEE SMC. He has serviced as a general chair, a program chair, and a committee member for many international conferences. He is the founder and a general chair of the series International Conferences on Wavelets Analysis and Pattern Recognition (ICWAPRs). He is the founder and chair of the Macau Branch of International Associate of Pattern Recognition (IAPR). He is an IEEE Fellow and IAPR Fellow.

Yang Lu (S'13) received the BSc degree in software engineering from the University of Macau, Macau, China, in 2012, where he is currently working towards the MS degree. His current research interests include pattern recognition, machine learning, hyperspectral image, and bioinformatics. He is a student member of the IEEE.

Huiwu Luo received the BS and MS degrees from Chongqing University, Chongqing, China, in 2008 and 2011, respectively. He is currently working towards the PhD degree in the Faculty of Science and Technology, University of Macau, Macau, China. His research interests include machine learning and image processing.

▷ **For more information on this or any other computing topic, please visit our Digital Library at** www.computer.org/publications/dlib.