Applied Mathematics & Information Sciences
*An International Journal*



# A Wavelet Transform Based Protein Sequence Similarity Model

*Jie Su and Junpeng Bao**

Department of Computer Science and Technology, Xi'an Jiaotong University, Xi'an 710049, P.R. China

Received: 11 Nov. 2012, Revised: 23 Jan. 2013, Accepted: 15 Feb. 2013
Published online: 1 May 2013

**Abstract:** Protein sequence analysis is an important tool for researchers to study on bio-informatics and molecular biology, such as proteins structure and function prediction, phylogenetic classification and different conservation pattern recognition. It is a significant open issue to quickly efficiently find the similar proteins from a large scale of protein repository. This paper proposes a new method based on Discrete Wavelet Transform (DWT) to measure the similarity of protein sequences, i.e. the ACDWT model, as well as two amino acid encoding methods (HPC and ADCC) according to hydropathy properties and dissociation constants respectively. The model employs only the approximation coefficients of DWT so that the feature vector is short. That brings the proposed model a great running time promotion. According to the phylogenic trees about nine ND5 proteins made from our model and others, the experimental results show that our model is efficient and a little better than the others.

**Keywords:** Protein Sequences Similarity, Discrete Wavelet Transform, Haar Wavelet

## 1. Introduction

As the rapid development of bio-informatics, a huge amount of protein sequence data presents a variety of challenges for bio-scientists. A lot of considerable efforts have been paying to find effective and reliable methods to deal with them. Rigden [14] presented that proteins with significant similar sequence are likely to have similar function. The key of protein sequence analysis is to detect the similarity of protein sequences.

There are various approaches to measure the similarity of protein sequences. The edit distance is a popular similarity measuring model [15]. It concerns with the minimum edit operations (insertion, deletion, or substitution) on individual amino acids so as to transform one sequence into another. This method depends on individual amino acids but involves in less protein features. Some methods divided amino acid sequences into different segments with scaled window or benchmark window, such as [1–5]. But the optimal length of the small peptide fragments is hard to know during division. In order to avoid this variable length problem, some methods measure the similarity of two sequences on global alignments, such as [6–13]. The basic steps of these methods are similar. Firstly, the protein sequences are translate into numeral signals; secondly, the numeral signals are converted into signal features; finally, the similarity of protein sequences is accessed by specific similarity of the signal features, for example, the Euclidean Distance or the cross-correlation coefficient. Zhang and Yu [5] map the protein primary sequence into a four-letter sequence, and then convert the sequence into 56 dimensional feature vectors by a complicated statistics and mathematical process.

In this paper, we present a protein sequences similarity model based on Discrete Wavelet Transform (ACDWT model), as well as two amino acid encoding methods according to hydropathy scale, pKa (acid dissociation constant) values of COOH and NH3+.

The rest of this paper is organized as follows. Section 2 introduces some related work on the problem. The similarity measuring model is described in the section 3. The experimental results are illustrated in the section 4. At last, conclusions and future work are stated in the section 5.

## 2. Related Work

According to the representation of a protein sequence, there are three types of the protein sequence analysis

* Corresponding author e-mail: baojp@mail.xjtu.edu.cn

methods: (1) digital signal based representation, (2) character sequence based representation, (3) graph based representation.

The digital signal based representation encodes a single amino acid into a number so that a protein sequence is converted into a digital signal sequence, such as [1–3, 10–12]. And then some digital signal analysis tools are used to extract the features of the protein sequence, for example, Empirical Model Decomposition (EMD), Discrete Fourier Transform (DFT) and Discrete Wavelet Transform (DWT). Wu et al. [1] believe that protein features are probably associated more with small peptide fragments than with individual amino acids. Thus a set of similar amino acids is assigned to a code, and a pattern is a sequence of codes with 4 numbers. Finally, the similarity of two protein sequences is measured according to the shared short patterns. The patterns are also used to cluster and predict the protein secondary structure. Wen et al. [2] apply DWT with various protein substitution models to find functional similarity of proteins. The full wavelet coefficients at a fixed level are used to measure the pair-wise similarity of protein sequences by a complex S function. Shi et al. [3] encode a protein sequence by an Electron-Ion Interaction Potential (EIIP) model, and decomposes the numeral signal by a Empirical Mode Decomposition so as to find functional similarity of proteins. They suggest a segmentation technique to handle long protein sequences and analyze the similarity of protein sequences by means of cross-correlation. Irena Cosic [10] represents a novel view on interactions of biomolecular, particular protein-protein and protein-DNA and assumes that these interactions are electromagnetic in their nature. A Resonant Recognition Model (RRM) is proposed to extract protein sequence linear information. It also encodes an amino acid by the EIIP value, but converts a coded sequence into the frequency domain by DFT. Seth and Duan [11] investigate the degree of overall similarity among protein sequence pairs of mouse proteome and examine the sequence similarity distributions at different levels of Gene Ontology (GO) tree. The overall similarity between two protein sequences is calculated by a model with the BLAST alignment scores. Chen et al. [12] present a decoy discrimination method by means of wavelet analysis. The signals of amino acid sequences and the profile of Solvent Accessibility Area from the protein conformation are de-noised by DWT. And a scoring function is developed to evaluate the similarity of the de-noised signals. The conformation that has the highest score is the most native-like one.

The character sequence based representation directly deals with the amino acid alphabet, such as [4–6, 8, 9, 13]. Kelil et al. [4] propose a Substitution Matching Similarity model to match amino acid subsequences. The model is especially designed for applications to non-aligned protein sequences. As a result, the first alignment-free algorithm, named CLUSS, was developed to cluster protein families. That has a great effect on both aligned and non-aligned protein families. Zhong et al. [5] explore protein sequence motifs with a K-means clustering algorithm and focus on characterizing the structural similarity in the sequence clusters so as to assess the clusters significance. Zhang and Yu [6] generate features from the hydropathy properties of amino acid sequences, and classify the 20 amino acids into four groups. A primary protein sequence is expressed by a four-letter sequence so that a full protein sequence is expressed by a 56 dimensional feature vector. Liu et al. [8] takes the entire sequence effect into account in order to avoid the variable length problem. Babaei et al. [9] introduce a novel method to derive protein networks based on their functional similarities. The method is employed to improve signature concordance and biological interpretability of breast cancer classification. Liu and Zhang [13] propose a novel method for phylogenetic analysis of H5N1 avian influenza virus. They use a four-letter sequence to express a primary protein sequence, and provide a curve mapping of virus protein sequence to construct a phylogenic tree without multiple alignments.

Yao et al. [7] introduce a graph based protein sequence representation. A protein dynamic 2D graph is drawn based on physicochemical properties of amino acids. It could reflect the innate structure of the protein sequence, rather than the apparent legitimate structure. Huang et al. [19] developed general knowledge by several data mining techniques, including decision tree, decision table and association rule algorithms to understand the protein stability change upon double mutation.

## 3. The ACDWT Model

### *3.1. Motivation*

As well-known, all proteins are made of 20 amino acids in different spatial structures and lengths. It is easy to express a protein by a sequence. But the question is that how to easily fast and accurately find the similar sequences in a large scale of protein repository. As mentioned above, the string based matching or searching methods can be directly employed, and they prefer probability theory to compute the proteins similarity. This approach is easy to find the straight supernatant duplicated similar pieces, but hard to find the latent similarity. For example, some sequences may have similar structure (or period, trend) but no identical string so that string based methods may have a poor performance. However, the digital signal analysis methods can deal with sequences in different spaces, including time domain and frequency domain. There are many good facilities to find the latent similarity of protein sequences.

In this paper, we also consider a protein sequence as a signal sequence, and employ the DWT to analyze and construct the feature vector of a protein sequence. Based

on the similarity matrix of the feature vectors, a set of protein sequences are clustered into different groups to build a phylogenic tree.

Our model is some similar to the Wen et al. [2] method, but we do not perform any complex function to compute the features. In fact, we only use a part of wavelet coefficients (i.e. the Approximate Coefficients) in the feature vector whereas the latter employs all of them (i.e. both Approximate Coefficients and Detail Coefficients). We believe that AC is enough to compute the protein similarity. So our model is called Approximate Coefficients of Discrete Wavelet Transform based Model (ACDWT Model). The most important advantage of the ACDWT Model is that it is simple, fast, fit for large scale of protein sequences.

Moreover, the paper introduces two protein sequence encoding methods (hydropathy properties code and acid dissociation constants code) according to the physicochemical properties of amino acids. These encoding methods are inspired by the substitution models, such as BLOUSUM [2] and EIIP[3].

Discrete Wavelet Transform. Fourier Transform (FT) and Wavelet Transform (WT) are two popular tools in the digital signal processing area. The FT can convert a signal sequence into the frequency domain so as to extract the frequency features, such as the energy focused frequencies. However, the FT assumes that the frequency spectrum covers the whole time axis from past to future. It implies the FT can not determine a specific frequency at an exact time point so that we only know a frequency feature happened but do not know when it happened. The WT is different from FT. The former can observe the features of a signal at both time domain and frequency domain so as to find some frequency feature at the specific time. The Discrete Wavelet Transform (DWT) deals with discrete data. It decomposes a signal into two parts: the approximation coefficients (AC) and the detail coefficients (DC). The AC expresses the whole global trend of the signal and holds most parts of the signal energy, whereas the DC expresses the local steep change details. Usually, the AC is half length of the signal. The AC at a level can be considered as the signal of the next level so that it can be decomposed recursively until only one number left. Hence, the DWT can observe the signal features at different scales. That is a great virtue to fast find the significant segments or focus on the right regions.

## *3.2. Encoding Methods*

### 3.2.1. Hydropathy properties code

A protein sequence is defined as a linear sequence of symbols from a 20-letter amino acid alphabet A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y. Zhang and Yu [6] presented a classified method that is based on the hydropathy scale including strongly hydrophilic (POL), strongly hydrophobic (HPO), weakly hydrophilic, weakly hydrophobic (Ambi), and others. The 20 amino acids are classified into four groups, according to the above hydropathy properties. Each group is assigned to a number, as shown in the Table 1. The 20 amino acids are also assigned to 20 integer numbers from 1 to 20, as shown in the Table 2. As a result, an amino acid is denoted by a 2-dimensions tuple, in which the first value is the code of the individual amino acid and the second value is the classifications code of the amino acid. For example, the amino acid Arg (R) is encoded to (1, 0), where 1 is the code of R and 0 is the group code of R. A protein sequence MNLFTS is encoded to(12, 1), (4, 0), (8, 1), (13, 1), (15, 2), (14, 2).

**Table 1** The group codes of amino acids

| Hydropathy Properties | Abbreviation | Amino acids | code |
|---|---|---|---|
| Strongly hydrophilic (polar) | POL | R,D,E,N,Q,K,H | 0 |
| Strongly hydrophobic | HPO | L,I,V,A,M,F | 1 |
| Weakly hydrophilic | Ambi | S,T,Y,W | 2 |
| Special | None | C,G,P | 3 |

### 3.2.2. Acid dissociation constants code

It is known that all amino acids contain the COOH and NH3+ groups. The two groups are weak acid groups that ionize in aqueous solutions. Their pKa values are 1.71–2.63 and 8.80–10.78, as shown in the Table 2. The capability of the groups donating and accepting electrons is essential to distinguish a proteins chemical properties. Moreover, the ionization of these groups has an important effect on a protein structure and catalytic activities of enzymes. As a result, the pKa value of the COOH and NH3+ groups can be used to encode a protein sequence. Thus, an amino acid is denoted by a 3-dimensions tuple, in which the first value is also the code of the individual amino acid, the second value is pKa value of COOH and the third value is pKa value of NH3+ .For example, the amino R is encoded to (1, 2.17, 9.04), where 1 is the code of R, 2.17is the pKa value of the COOH in the R and 9.04 is the pKa value of NH3+ in the R. The protein sequence MNLFTS is encoded to (12, 2.28, 9.21), (4, 2.02, 8.80), (8, 2.36, 9.60), (13, 1.83, 9.13), (15, 2.63, 10.43), (14, 2.21, 9.15).

### 3.2.3. Similarity Metric

Wen el al. [2] used both approximation coefficient and detail coefficient to calculate the similarity of two protein

**Table 2** The acid dissociation constants[7] and the code of individual amino acid

| Amino acid | code | pKa-COOH | pKa-NH3+ |
|---|---|---|---|
| R | 1 | 2.17 | 9.04 |
| D | 2 | 2.09 | 9.82 |
| E | 3 | 2.19 | 9.67 |
| N | 4 | 2.02 | 8.80 |
| Q | 5 | 2.17 | 9.13 |
| K | 6 | 2.18 | 8.95 |
| H | 7 | 1.82 | 9.17 |
| L | 8 | 2.36 | 9.60 |
| I | 9 | 2.36 | 9.68 |
| V | 10 | 2.32 | 9.62 |
| A | 11 | 2.34 | 9.69 |
| M | 12 | 2.28 | 9.21 |
| F | 13 | 1.83 | 9.13 |
| S | 14 | 2.21 | 9.15 |
| T | 15 | 2.63 | 10.43 |
| Y | 16 | 2.20 | 9.11 |
| W | 17 | 2.38 | 9.39 |
| C | 18 | 1.71 | 10.78 |
| G | 19 | 2.34 | 9.60 |
| P | 20 | 1.99 | 10.60 |

sequences. Many researches have suggested that the AC part expresses the whole global trend of the signal and holds most portion of the signal energy. The DC part can supply some supplement information. If we prefer to observe the exact details of a piece of protein, it is better to check the protein sequence character by character. Obviously, it is not fit for fast efficiently computing of the protein similarity in a large scale repository. In fact, the DC can be ignored as the redundant part when we do not expect an over-refined similarity model. Consequently, the paper proposes the Approximate Coefficients of Discrete Wavelet Transform based Model (ACDWT Model) to express the features of a protein sequence only by AC, rather than full coefficients of DWT. A direct benefit of the ACDWT Model is that the length of AC vector is half of the whole coefficients vector. That definitely leads to a promotion of running time.

The following is the pseudo code of the ACDWT Model, in which the encoding method is the hydropathy properties code method (HPC). Alternatively, the acid dissociation constants code method (ADCC) can be employed at the step c) and d). When the protein sequence encoding method is changed, the ACDWT Model may produce a different similarity value, i.e. the similarity matrix is different.

*The ACDWT Model Algorithm*

*Input: Protein Sequences $S_1$, $S_2$, and the decomposition level M.*

*Output: The similarities of $S_1$ and $S_2$ at level M.*

*a) Read protein sequences $S_1$.*

*b) Read protein sequences $S_2$.*

*c) Get $S_1$'s encoding sequence $L_{hp1}$*

*d) Get $S_2$'s encoding sequence $L_{hp2}$,*

*e) Get the Signal Length n,*

$$n = max(|L_{hp1}|, |L_{hp2}|) \tag{1}$$

*f) Validate the decomposition level M,*

$$M = min(M, \lceil log_2(n) \rceil) \tag{2}$$

*g) Execute the Discrete Wavelet Transform(DWT) with the Haar Wavelet on $L_{hp1}$ at level M to get the $L_{hp1}$'s approximation coefficients $AC_1$,*

$$(AC_1, DC_1) = DWT(L_{hp1}, Haar, M) \tag{3}$$

*h) Execute the Discrete Wavelet Transform(DWT) with the Haar Wavelet on $L_{hp2}$ at level M to get the $L_{hp2}$'s approximation coefficients $AC_2$,*

$$(AC_2, DC_2) = DWT(L_{hp2}, Haar, M) \tag{4}$$

*i) Calculate the value of Cosine between $AC_1$ and $AC_2$, which is the similarity value of the protein sequences $S_1$, $S_2$ at level M,*

$$Sim(S_1, S_2, M) = \frac{AC_1 \cdot AC_2}{|AC_1| \times |AC_2|} \tag{5}$$

*j) End*

where $|L_{hp1}|$ denotes the length of the signal $L_{hp1}$. The length of DWT approximation coefficients (AC) shrunk to half while the level goes up a step, i.e. the length of AC at the level j is,

$$|AC^j| = \frac{1}{2^j} n (j = 1, 2, ..., k) \tag{6}$$

where $n$ is the length of the original signal, $j$ is DWT decomposition level, $k$ is the max level.

The larger values of $Sim(S_1, S_2, M)$, the more similarity of the two proteins $S_1$ and $S_2$. Consequently, a symmetric matrix $D$ is constructed based on the $Sim$ value. Then, we can execute the Slink clustering algorithm to build a phylogenic tree based on the similarity matrix in order to clearly demonstrate the model's result.

## 4. Experimental Results

The phylogenetic trees. Nine ND5 (NADH dehydrogenase subunit5) proteins in the NCBI database are used in the experiments. The dataset includes human (Homo sapiens, AP-000649), gorilla (Gorilla gorilla, NP-008222), common chimpanzee (Pantroglodytes, NP-008196), pigmy chimpanzee (Pan paniscus, NP-008209), fin whale (Balenoptera physalus, NP-006899), blue whale (Balenoptera musculus, NP-007066), rat (Rattus norvegicus, AP-004902), mouse (Mus musculus, NP-904338), and opossum (Didelphis virginiana, NP-007105). The dataset is also used by Zhang [6] and Yao [7]. The ACDWT Model is compared with their models.

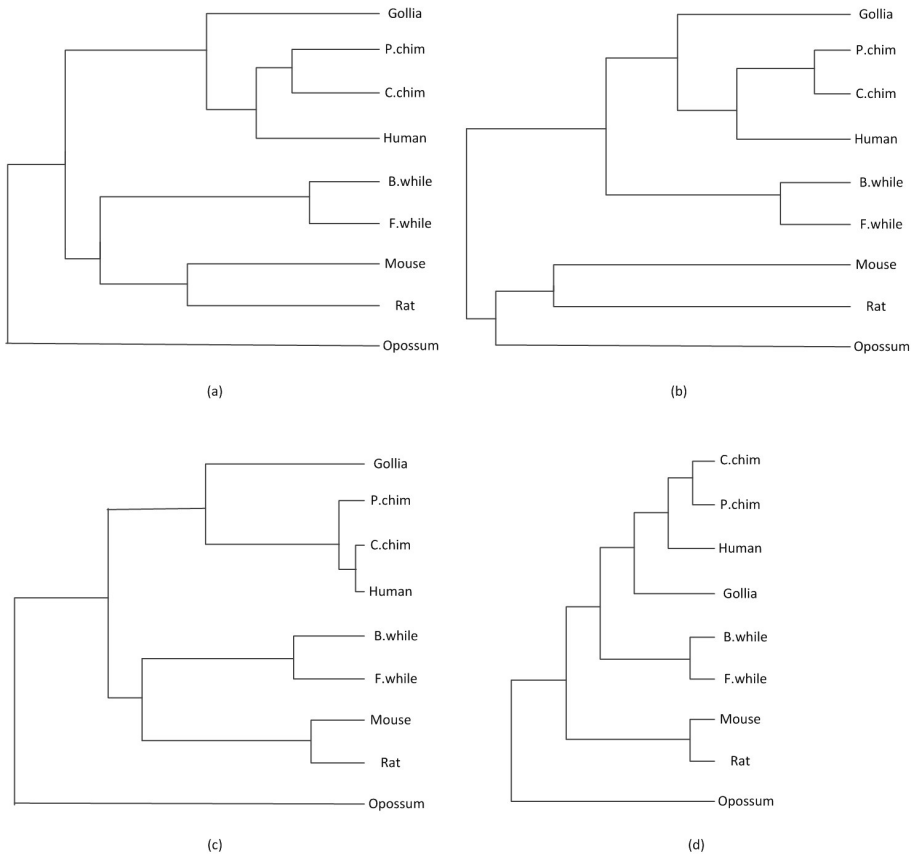


**Figure 1** The phylogenetic tree constructed by ACDWT Model (a), Zhang's Model[6] (b), Yao's Model[7] (c), and MEGA software (d).

The Fig. 1(a) is the phylogenetic tree built by the ACDWT Model. Though the two encoding methods produce different similarity matrix, they result in the same phylogenetic tree. The Fig. 1(a) shows that the ND5 proteins of human, common chimpanzee, pigmy chimpanzee and gorilla are similar to each other, the group of fin whale and blue whale is similar to the group of rat and mouse, rather than the group of human and chimpanzee. On the other hand, the protein of opossum (the most remote species from the other mammals) is most dissimilar to the others among the nine species.

The Fig. 1(b) is the phylogenetic tree constructed by Zhang's model [6]. It is very like the Fig. 1(a). But it suggests that the group of rat and mouse is more similar to opossum, rather than group of whales. The Table 3 ranks the similarity between human and other species. The difference of encoding methods has no effect on the ACDWT Models final result. According to the ACDWT Model, gorilla is more similar to human than whale whereas Zhang [6] suggests that whale is closer to human than gorilla.

We also illustrate the phylogenetic tree built by Yao [7] in the Fig. 1(c) and one created by MEGA software [6] in the Fig. 1(d). It is clear that the Yao's model [7] is the worst result, and no one is identical to the MEGA's result. Whatsoever, ACDWT Model's result is more consistent to the known fact of evolution [16–18]. As a result, it is a fact that the precision of ACDWT Model is a little better than Zhang [6] and Yao [7] on this dataset.

**Table 3** The similarity between human and other species

| Model | Similarity Rank |
|---|---|
| ACDWT Based on HPC | P.Chim >C.Chim >Gorilla >F.whale >B.whale >Mouse >Rat >Opossum |
| ACDWT Based on ADCC | P.Chim >C.Chim >Gorilla >F.whale >B.whale >Mouse >Rat >Opossum |
| Zhang's Model | P.Chim >C.Chim >B.whale >F.whale >Gorilla >Mouse >Rat >Opossum |

The improvement of efficiency. Moreover, the ACDWT Model has a shorter feature vector than Wen's model [2]. The ACDWT Model only uses the approximation coefficients of DWT, whereas the latter employs both the approximation coefficients and the detail coefficients. The calculation process of the

ACDWT Model is simple so that it is easy to understand. Consequently, the ACDWT Model has a great advantage of the running time promotion. Though the similarity matrix of the two models are different, the clustering results of them are identical. The Fig. 2 compares the running times of the ACDWT Model with the Wen's model [2]. It is clear that the running time of ACDWT Model is less than that of Wens model. In the best case, the former is only 64.14% of the latter in our test.

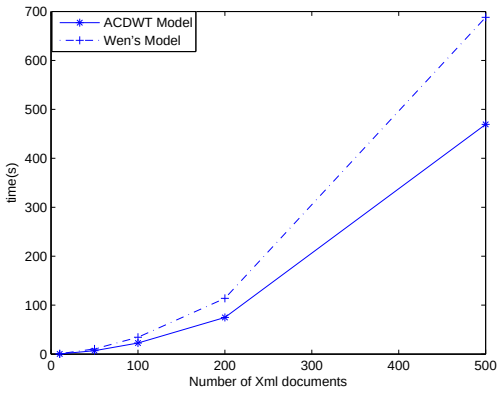


**Figure 2** The running time in seconds of the ACDWT Model and Wen's Model

The optimal wavelet. As well known, the Haar Wavelet is the simplest DWT, but it may not be the best wavelet to measure the protein similarity. We test 50 wavelets in order to find a more efficient wavelet, including haar, bior1.1, bior1.3, bior1.5, bior2.2, bior2.4, bior2.6, bior2.8, bior3.1, bior3.3, bior3.5, bior3.7, bior3.9, bior4.4, bior5.5, bior6.8, db2-db20, sym1-sym10, coif1-coif10.

In order to compare the performance of various wavelets, we define a distance error to measure the consistence of them. It is assumed that the original encoded protein sequence contains full information so that the direct distances between the original sequences are the basic coordinates. Obviously, this is a low efficient process because the original sequence is much too long. The ACDWT model exploits only the approximate coefficients of the transformed signal, which keep the main trend of the original full length signal. It is expected that the distances based on the short vectors of approximate coefficients keep the same trend with the distances based on the original one. Namely, the difference between the two distances matrix should be small. Hence, a distance error is defined in the Equation 7 to measure the standard deviation of the difference between the original protein encoding sequence distance and the other distance based on the ACDWT model.

$$DE = \sqrt{\frac{\sum_{i=1}^{P}(\Delta d_i - \overline{\Delta d})^2}{P}} \tag{7}$$

Where, $P$ is the number of signal pairs. $\Delta d_i = ds_i - dc_i$ denotes the difference between the original signal distance of the ith protein encoding sequence pair and the ACDWT model's distance of that pair. The $\overline{\Delta d}$ is the average of $\Delta d_i$. If $\Delta d_i$ at each point is identical, then the distance error is 0. It means that the distributions of two distances are the same.

The Fig. 3 shows that the bior wavelets are better other wavelets in the ACDWT model no matter based on HPC or ADCC, in which the bior3.1, bior3.3 and bior3.5 (the 9th, 10th and 11th) are the best. They have almost the same distance error in both curves. The bior3.5 has the minimum distance error in the ACDWT model based on HPC whereas the bior3.1 has the minimum value in the ADCC case.

The Fig. 4 compares the running time of both HPC and ADCC on 50 wavelets over 500 proteins. It shows that Haar is the fastest wavelet, the sym10 and db20 are the slowest, the db11- db19 and sym9 are much slow wavelets, the others are similar and close to Haar. It suggests that the Haar is the simplest and fastest wavelet whereas its performance is on the average level. The bior wavelets, especially the bior3.1, bior3.3 and bior3.5 are a litter slower than Haar, but they apparently outperform the others. The sym10 and db20 are the worst, which are too slow to deal with the large scale data.

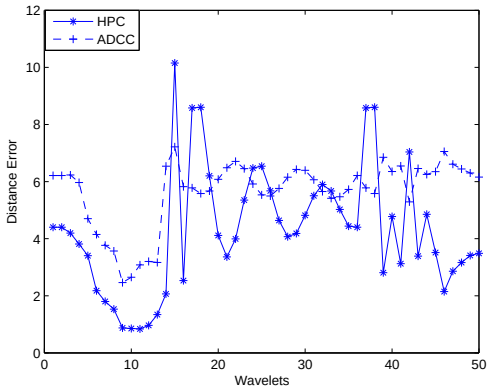


**Figure 3** The distance error of the ACDWT model with HPC and ADCC on 50 wavelets. The wavelets from 1 to 50 are haar, bior1.1, bior1.3, bior1.5, bior2.2, bior2.4, bior2.6, bior2.8, bior3.1, bior3.3, bior3.5, bior3.7, bior3.9, bior4.4, bior5.5, bior6.8, db2-db20, sym1-sym10, coif1-coif10.

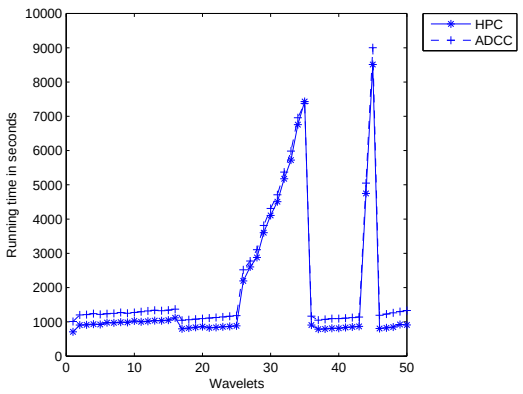


**Figure 4** The running time of the ACDWT model with HPC and ADCC on 50 wavelets over 500 proteins. The wavelets from 1 to 50 are haar, bior1.1, bior1.3, bior1.5, bior2.2, bior2.4, bior2.6, bior2.8, bior3.1, bior3.3, bior3.5, bior3.7, bior3.9, bior4.4, bior5.5, bior6.8, db2-db20, sym1-sym10, coif1-coif10.

## 5. Conclusions and Future Work

In order to analyze a huge amount of protein sequence data effectively and precisely, the paper introduces a new model to measure the similarity of protein sequences based on discrete wavelet transform. The main contributions are:

(1) Two amino acid encoding methods (HPC and ADCC) are proposed according to hydropathy properties and dissociation constants respectively. It is helpful to conserve the important physicochemical properties of amino acids.

(2) A simple DWT based model, i.e. the ACDWT model, is suggested to get the feature vectors of protein sequence and measure their similarity. The model employs only the approximation coefficients of DWT, ignores the detail coefficients so as to shorten the feature vector. Since the AC part expresses the whole global trend of the signal and holds most energy of the signal. We believe it is enough to deal with the similarity of proteins.

(3) The bior wavelets are better than others, especially the bior3.1, bior3.3 and bior3.5.

The Experimental results support our view. The performance of the ACDWT model is better than that of Zhangs model and Yaos model.

However, the structure information of protein molecule is crucial to finally evaluate the similarity of the proteins. We are going to research structure-included proteins measure model to solve the issue.

## Acknowledgement

This research is supported by National Natural Science Foundation of China (Grant 60903123), the Fundamental Research Funds for the Central Universities and the Baidu Theme Research Plan on Large Scale Machine Learning and Data Mining.

## References

[1] K.-P. Wu, et al., A New Similarity Measure among Protein Sequences. Proc. the IEEE Bioinformatics Conference, pp. 347-352 (2003)

[2] Z.-N. Wen, et al., Analyzing Functional Similarity of Protein Sequences with Discrete Wavelet Transform. Computational Biology and Chemistry. **29(3)**: 220-228 (2005)

[3] F. Shi, Q.J. Chen, X.H. Niu, Functional Similarity Analyzing of Protein Sequences with Empirical Mode Decomposition. Proc. the Fourth International Conference on Fuzzy Systems and Knowledge Discovery, pp. 766-770 (2007)

[4] A. Kelil, et al., CLUSS: Clustering of Protein Sequences Based on a New Similarity Measure. BMC Bioinformatics, **8**: 286-305 (2007)

[5] W. Zhong, et al., Mining Protein Sequence Motifs Representing Common 3D Structures. Proc. IEEE Conference on Computational Systems Bioinformatics, pp. 215-216 (2005)

[6] Y.S. Zhang, X.T. Yu, Analysis of Protein Sequence Similarity. Proc. the IEEE Fifth International Conference on Bio-Inspired Computing: Theories and Applications, pp. 1255-1258 (2010)

[7] Y.H. Yao, et al., Analysis of Similarity/Dissimilarity of Protein Sequences. Proteins. **73(4)**: 864- 871 (2008)

[8] H. Liu, X. Liu, Y. Yao, Identification of Secretory Proteins Based on Similarity of Amino Acid Sequences. Proc. the third International Conference on Biomedical Engineering and Informatics, pp. 2316-2320 (2010)

[9] S. Babaei, et al., Integrating Protein Family Sequence Similarities with Gene Expression to Find Signature Gene Networks in Breast Cancer Metastasis. Lecture Notes in Computer Science, Vol. **7036**, pp. 247-259 (2011)

[10] I. Cosic, molecular Bioactivity: Is It Resonant Interaction Between Macromolecules? Theory and Applications. IEEE Transactions on Biomedical Engineering, **41(12)**: 1101-1114 (1994)

[11] P. Seth, Z.-H. Duan, Study of the Relationship between Mus musculus Protein Sequences and their Biological Functions. Proc. WASE International Conference on Information Engineering, pp. 557-560 (2009)

[12] M.X. Chen, et al., Wavelet Transform Based Protein Decoy Discrimination. Proc. the third International Conference on Bioinformatics and Biomedical Engineering, pp. 1-4 (2009)

[13] Y.Q. Liu, Y.S. Zhang, A New Method for Analyzing H5N1 Avian InfluenzaVirus. Journal of Mathematical Chemistry. **47(3)**: 1129-1144 (2010)

[14] D. Rigden, From protein structure to function with bioinformatics. Springer, Heidelberg, 2009

[15] V.I. Levenstein. Binary codes capable of correcting insertions and reversals. Sov. Phys. Dokl. **10**: 707-710 (1966)

[16] M. Li, et al., An information-based sequence distance and its application to whole mitochondrial genome phylogeny. Bioinformatics. **17(2)**: 149-154 (2001)

[17] H.H. Otu, K. Sayood, A new sequence distance measure for phylogenetic tree construction. Bioinformatics. 19(16):2122- 2130(2003)
[18] V. Makarenkov, F. Lapointe, A weighted least-squares approach for inferring phylogenies from incomplete distance matrices. Bioinformatics. **20(13)**: 2113-2121 (2004)
[19] L.-T. Huang, et al., Data mining application in biomedical informatics for probing into protein stability upon double mutation. Applied Mathematics and Information Sciences, **7(2S)**: 563S-570S (2013)

**Jie Su** is a master student of Computer Science in Xi'an Jiaotong University. She has been studying on Data Mining and Machine Learning for 3 years, especially about algorithms of mining useful information from semi-structural data and sequence data. She proposed the Wavelet Transform Based Structural Similarity Model and the Level Edit Distance Model to detect the structural similarity of XML texts, Homepages of high schools, UML models, Protein sequences and other massive data. She has published 3 academic papers and submitted 1 invention patent.

**Junpeng Bao** is an Associate Professor of Computer Science in Xi'an Jiaotong University. He received his Computer Science PhD from Xi'an Jiaotong University (China) in 2004. His current research interests are Data Mining, Machine Learning, and Artificial Intelligence. He had visited University of Hertfordshire in UK for one year as a visiting scholar supported by the Royal Society. He possessed 3 China invention patents, published 1 book (A Guide to Artificial Intelligence) and 17 academic papers.