Available online at www.sciencedirect.com
ScienceDirect
Pattern Recognition Letters 29 (2008) 616–628
Pattern Recognition Letters
www.elsevier.com/locate/patrec

An efficient conserved region detection method for multiple protein sequences using principal component analysis and wavelet transform

Chieh-Yuan Tsai *, Chuang-Cheng Chiu

Department of Industrial Engineering and Management, Yuan Ze University, 135 Yuantung Road, Chungli City, Taoyuan County 320, Taiwan

Received 13 June 2006; received in revised form 13 July 2007
Available online 8 December 2007

Communicated by M. Singh

Abstract

This paper proposes an efficient conserved region detection method for multiple protein sequences. Instead of detecting conserved regions directly from the set of all participatory protein sequences, the proposed method separates the detection process as two stages. In the fist stage, a serial of principal component analysis (PCA) techniques are applied to infer the common ancestor protein from the participatory proteins based on a hypothetical evolutionary history. Then, wavelet transform is employed to derive conserved regions from the common ancestor protein in the second stage. The detected conserved regions are considered as the common conserved regions of the original protein sequences. A set of experiments indicate that the two stage strategy makes the proposed method not only prevents the residue divergence problem but also increases the detection accuracy and efficiency.

Keywords: Protein sequence analysis; Conserved region detection; Principal component analysis; Wavelet transform

1. Introduction

A conserved region in a protein sequence is the region composed of a number of successive amino acids that carries specific functional importance for the protein sequence (Orengo et al., 2003). During the molecular evolution, the amino acids within a conserved region appear to resist mutations, whereas others tend to make diverse substitutions. The characteristic of a conserved region causes the ancestor protein remain important functions to its children proteins. The conserved regions all appeared in a set of proteins sequences can be considered as the common patterns for theses sequences. The patterns are tremendously useful for biologists in protein analysis tasks, such as unknown protein classification, protein structure identification, protein family annotation, and so on. Therefore, how to detect the conserved regions from a set of protein sequences has been one of the major focuses in the fields of bioinformatics.

Traditional conserved region detection methods are divided into five main categories, including symbol frequency, symbol entropy, stereochemical property, mutation data and weighted sequence methods. Before performing these methods, all protein sequences are preprocessed by a multiple sequence alignment (MSA) task. MSA makes all aligned protein sequences have the same length for acquiring further protein homology information (Chan et al., 1992). Then, based on the scoring principle of the adopted method, the conservation score of each aligned column is calculated by involving all participatory proteins. The conservation score of an aligned column represents the degree that the column falls into a conserved region. The higher the conservation score for the aligned column, the higher the possibility that the aligned column belongs to a conserved region.

Symbol frequency-based methods (Wu and Kabat, 1970; Jores et al., 1990; Lockless and Ranganathan, 1999) consider amino acids as symbols in a uniformly

* Corresponding author. Tel.: +886 3 4638800x2512; fax: +886 3 4638907.
E-mail address: cytsai@saturn.yzu.edu.tw (C.-Y. Tsai).

0167-8655/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2007.11.013

diverse alphabet and focus on the relative frequency of these symbols, but do not account for sequence redundancy in the alignment. Symbol entropy-based methods (Shannon, 1948; Shenkin et al., 1991; Gerstein and Altman, 1995) which are the specialization of symbol frequency-based methods, consider the relative frequencies of symbols using Shannon's entropy or its variations. Stereochemical property-based methods (Williamson, 1995; Mirny and Shakhnovich, 1999) identify conserved regions from the stereochemical properties of the amino acids in each aligned column of the alignment. Mutation data-based methods (Karlin and Brocchieri, 1996; Lichtarge et al., 1996; Thompson et al., 1997; Armon et al., 2001) quantify stereochemical variability in each aligned column of the alignment using the mutation data of a specific substitution matrix, such as BLOSUM62, PAM120, and so on. Weighted sequence-based methods (Sander and Schneider, 1991; Landgraf et al., 1999) not only detect conserved regions, but also implement multiple sequence alignment. Except the above methods, Valdar (2002) proposed a synthetic conserved region detection method which benefits from the advantages of symbol entropy and stereochemical property-based methods. Thus, whether an aligned column is included within the conserved region or not depends on its symbol diversity, stereochemical diversity, and gappiness. Although these methods are diverse in their scoring principles and approaches, they share a common characteristic: the evaluation of the conservation score needs to take all participatory proteins into account simultaneously. As the number of participatory protein sequences increases, however, residue divergence in each aligned column grows dramatically. The problem becomes worse especially when the participatory protein sequences share less than 25% sequence identity (Rost, 1999).

To avoid this problem, a novel conserved region detection method is proposed in this paper. Instead of detecting conserved regions directly from the set of all participatory protein sequences, the proposed method separates the detection process into two stages. The fist stage is to infer the common ancestor protein of the participatory proteins based on a reasonable molecular evolutionary relationship, then deriving the conserved regions from the common ancestor protein in the second stage. This is based on the truth that the conserved regions appeared on the ancestor protein are similar to the ones appeared on its children proteins. For that reason, the detection result based on the common ancestor protein theoretically is close to the result based on its children proteins. Therefore, conducting conserved region detection for a single protein, i.e. the common ancestor protein, not only prevents the residue divergence problem but also increases the detection accuracy and efficiency.

Furthermore, in a protein the amino acids within a conserved region reveal stronger functional responses than other amino acids. The functional response of each amino acid is a numerical value which can be measured from physical or chemical experiments. For example, electron–ion interaction potential (EIIP) is one of common functional responses which describe the energy of the delocalized electrons of each amino acid (Cosic, 1994). By introducing the EIIP values of all amino acids, the expression format for each protein can be represented as a numerical vector composed of amino acid's functional responses instead of using the original string sequence composed of amino acid's alphabets, which is another important feature of the proposed method. Besides considering the characteristic of conserved regions, the conserved region can be quickly detected through this transformation because the sections in a numerical vector with higher functional responses can be identified easily.

The rest of this paper is organized as follows. The overall process of the proposed conserved region detection method is gone into particulars in Section 2. Section 3 analyzes the performance of the proposed method using a number of benchmark datasets. The detection results are then compared with a popular conserved region detection tool to show the merit of the proposed method. Section 4 evidences the usage feasibility of the proposed method in practice. Finally, a summary and conclusion are presented in Section 5.

2. The proposed conserved region detection method

This section presents an efficient method to detect conserved regions for multiple protein sequences. The proposed method consists of four major processes including multiple sequence alignment, string–vector transform, ancestor inference, and conserved region detection. Fig. 1 illustrates the framework of the proposed method where the four major processes are highlighted with shaded color.[1] To clearly explain the proposed method, five protein sequences, 1ABOA, 1YCSB, 1PHT, 1HVA and 1BB9, acquired from the dataset "1aboA_ref1-reference 1" in the benchmark alignment database BAliBASE (http://www-igbmc.u-strasbg.fr/BioInfo/BAliBASE2), serves as the example sequences in the following introduction. The details about the BAliBASE database can be referred to Thompson et al. (1999).

2.1. Multiple sequence alignment

Similar to traditional conserved region detection methods, the first process of the proposed method is to conduct multiple sequence alignment (MSA), so that all original protein sequences are aligned as equal-length sequences. To fulfill the need, a popular global MSA tool, called ClustalW (http://www.ebi.ac.uk/clustalw), is adopted in this research. In ClustalW, the pairwise similarity scores for these original protein sequences are first calculated using a fast approximate approach (Bashford et al., 1987). Then,

[1] For interpretation of color in Fig. 1, the reader is referred to the web version of this article.

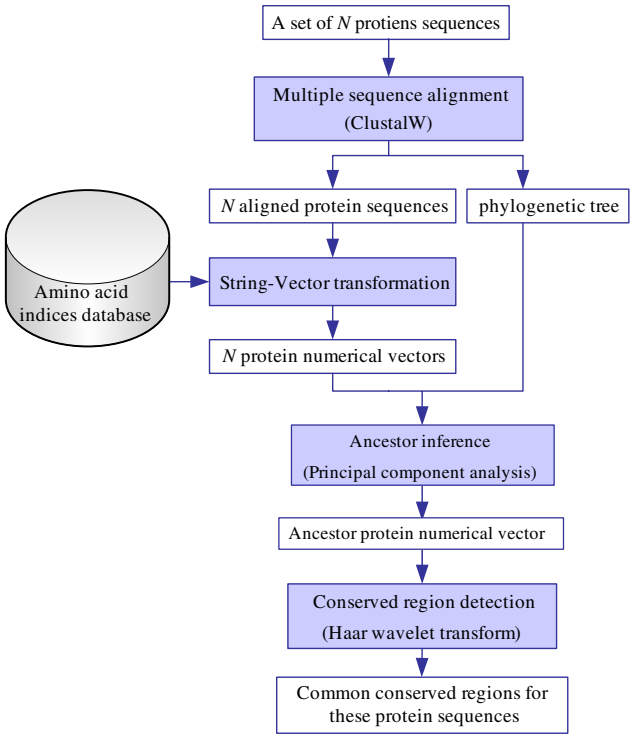

Fig. 1. The framework of the proposed conserved region detection method.

an unrooted tree is constructed using the neighbor-joining approach based on the calculated pairwise similarity scores (Saitou and Nei, 1987). Subsequently, the root of the unrooted tree is placed by the mid-point manner (Thompson et al., 1994b), so that the rooted tree is considered as a phylogenetic tree used to direct the progressive alignment for all protein sequences. Note that the rooted phylogenetic tree describes the evolutionary history of these protein sequences and will serve as the input in the third process of the proposed method, ancestor inference. The detail description of ClustalW algorithm can be found in (Thompson et al., 1994a). After manipulating ClustalW, the alignment result of the five example protein sequences, as well as their phylogenetic tree, are shown in Fig. 2. Note that a gap in an aligned protein sequence is represented as "-".

2.2. String–vector transformation

As mentioned above, in a protein the amino acids within a conserved region reveal stronger functional responses than other amino acids. The functional response of each amino acid can be measured from physical or chemical experiments and represented by a numerical value. For example, electron–ion interaction potential (EIIP), describing the energy of the delocalized electrons of each amino acid, is one of the most common functional responses. When each amino acid is descried by its corresponding EIIP value, therefore, all aligned protein sequences represented by alphabet string format can be transformed into to the numerical vectors consisting of a series of functional response values. Fig. 3 illustrates that the five example aligned protein sequences are represented by five numerical protein vectors according to EIIP. Note that the functional responses of amino acids can be referred to AAindex database (http://www.genome.ad.jp/dbget/aaindex.html) which stores up to 516 response values of amino acids (Cosic, 1994), and the functional responses of a gap represented as "-" are substituted by zeros.

2.3. Ancestor inference

Through string–vector transformation, an original protein sequence is represented as a vector $\boldsymbol{x}_i = [x_{i1}, \cdots, x_{ij}, \cdots, x_{in}]^T$ where x_{ij} is the functional response of the jth

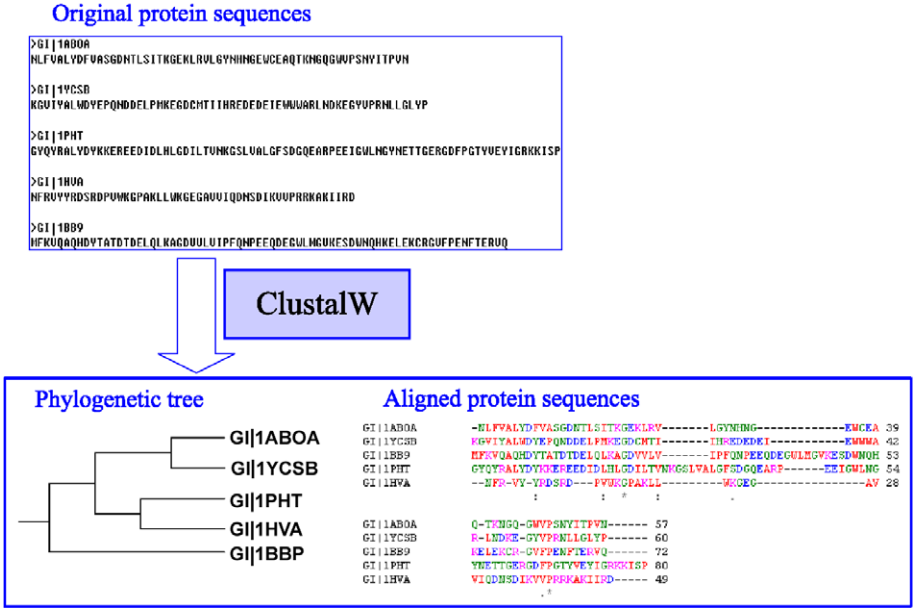

Fig. 2. The alignment result of the five example protein sequences using ClustalW.

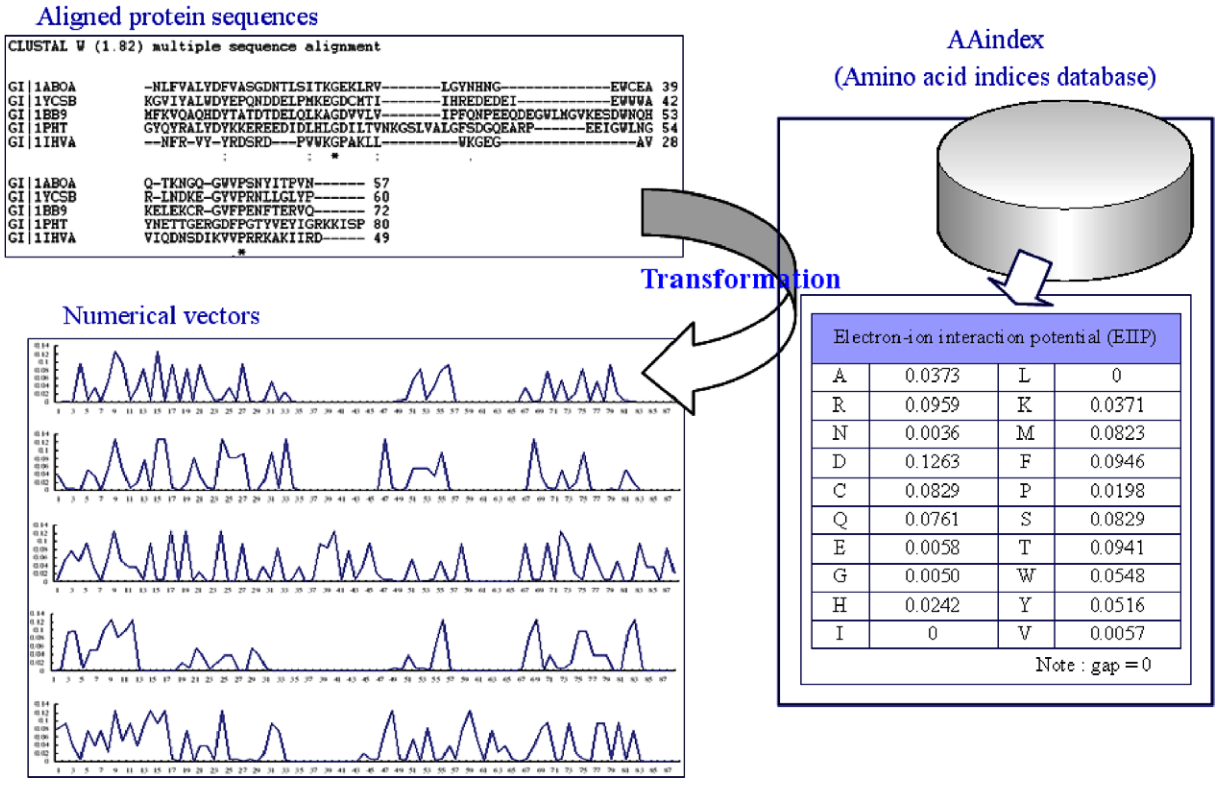

Fig. 3. The string–vector transformation process of the five aligned protein sequences.

residue of the ith protein vector and n is the number of residues. In one molecular evolution during overall evolutionary history, assume a parent protein vector $\boldsymbol{y} = [y_1, \cdots, y_j, \cdots, y_n]^T$ generates m children protein vectors $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_i, \cdots, \boldsymbol{x}_m$. Therefore, the process of this molecular evolution among $\boldsymbol{y}$ and $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_i, \cdots, \boldsymbol{x}_m$ can be represented as

$$\begin{cases} \boldsymbol{x}_1 = w_1 \times \boldsymbol{y} \\ \vdots \\ \boldsymbol{x}_m = w_m \times \boldsymbol{y} \end{cases} \Rightarrow [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_i, \cdots, \boldsymbol{x}_m] = \boldsymbol{y} \times [w_1, \cdots, w_i, \cdots w_m] = \boldsymbol{y} \times \boldsymbol{w} \tag{1}$$

where $\boldsymbol{w} = [w_1, \cdots, w_i, \cdots, w_m]$ is a m-dimension vector in which the ith element w_i represents a scale used to translate the parent protein vector $\boldsymbol{y}$ into the ith children protein vector $\boldsymbol{x}_i$. If any a children protein vector $\boldsymbol{x}_i$ and its corresponding translation scale w_i are known for all i, the parent protein vector $\boldsymbol{y}$ can be inferred by the Eq. (2), which is derived from Eq. (1).

$$\begin{cases} \boldsymbol{y} = 1/w_1 \times \boldsymbol{x}_1 \quad \text{or} \\ \vdots \\ \boldsymbol{y} = 1/w_{m-1} \times \boldsymbol{x}_{m-1} \quad \text{or} \\ \boldsymbol{y} = 1/w_m \times \boldsymbol{x}_m \end{cases} \tag{2}$$

Unfortunately, the translation vector $\boldsymbol{w}$ is intrinsically unknown. To infer a meaningful parent protein vector $\boldsymbol{y}$ under the source-blind situation, this paper applies principal component analysis (PCA) technique (Jolliffe, 1986) to estimate a vector $\boldsymbol{w}' = [\frac{1}{w_1}, \cdots, \frac{1}{w_i}, \cdots, \frac{1}{w_m}]$ from the m children protein vectors reasonably. PCA is an important and essential technique for data reduction, image compression, and feature extraction (Fortuna and Capson, 2004; Guru and Punitha, 2004). Let an $n \times m$ matrix $\boldsymbol{D}$ record all m children protein vectors of lengths are n equally. That is, each of m columns in $\boldsymbol{D}$ represents a children protein vector, whereas the n rows in $\boldsymbol{D}$ represent the n residue elements of m children protein vectors. When each column in $\boldsymbol{D}$ has standardized to have zero mean and unit variance, the covariance matrix of $\boldsymbol{D}$ equals to $\boldsymbol{D}^T\boldsymbol{D}$. By implementing PCA on $\boldsymbol{D}^T\boldsymbol{D}$, the eigenvector $\boldsymbol{v}_1$ corresponding to maximum eigenvalue λ_1 among m eigenvalues is retrieved to represent $\boldsymbol{w}'$. The derived eigenvector $\boldsymbol{v}_1$ ($=\boldsymbol{w}'$) makes the parent protein vector $\boldsymbol{y}$, i.e. the first principal component, be capable of distinguishing its m children protein vectors as well as possible. Consequently, the parent protein vector $\boldsymbol{y}$ can be calculated by Eq. (2).

However, applying PCA to infer the common ancestor protein vector directly from all participant protein vectors could be problematic if the molecular evolutionary relationships among these proteins are not considered. In the proposed method, therefore, the common ancestor protein inference is guided by the phylogenetic tree generated by ClustalW. The phylogenetic tree provides a hypothetical evolutionary history depicting how the current proteins (leaf nodes of the tree) are diverged from hypothesized proteins (internal nodes of the tree), and how the hypothesized proteins are evolved from the common ancestor protein (root node of the tree) during the molecular evolution. By following the branching order in the tree, the common ancestor protein vector at the top level of the tree can be inferred progressively from the participatory protein vectors at the bottom level by applying a serial of PCA approaches. Fig. 4 illustrates how a serial of PCA approaches are applied to the five example protein sequences by following the hypothetical evolutionary history described in the phylogenetic tree. It shows that the hypothesized protein vector A is the parent of two protein vectors 1ABOA and 1YCSB. Therefore, A can be inferred from 1ABOA and 1YCSB through the proposed PCA approach. Fig. 5 illustrates visually the inference process for the hypothesized protein vector A. Similarly, the hypothesized protein vector B is inferred from its two protein vectors 1PHT and 1HVA through the proposed PCA approach, and then the hypothesized protein vector C is inferred from these tow inferred protein vectors A and B. Finally, the common ancestor protein vector D is inferred from the C and 1BB9.

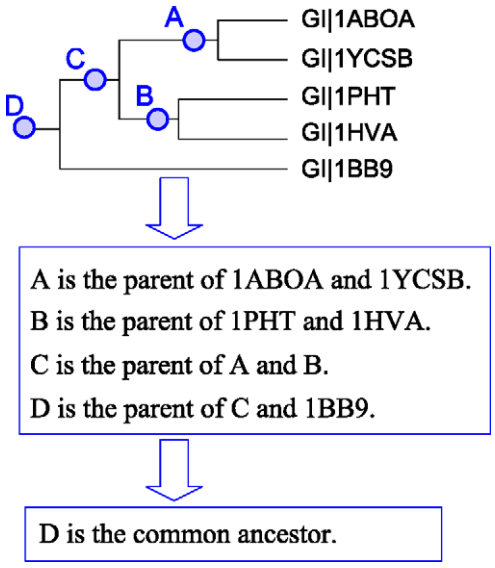

Fig. 4. The hypothetical evolutionary history for the five example proteins.

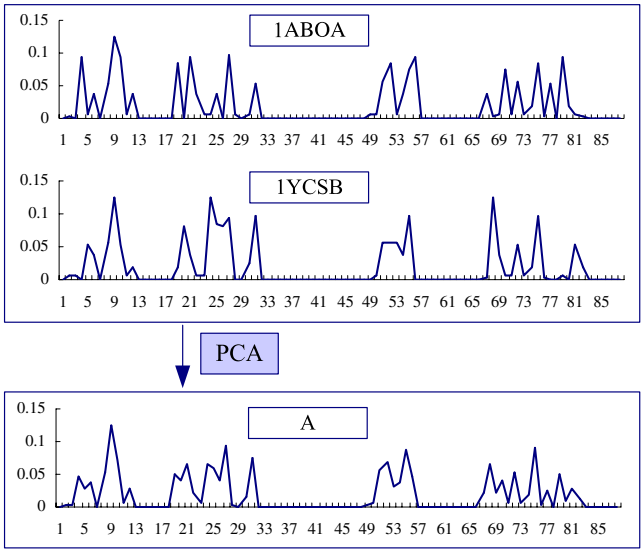

Fig. 5. Inferring the hypothesized protein vector A from its two children protein vectors using PCA approach.

2.4. Conserved region detection

After completing the process of ancestor inference, the common ancestor protein represented as a numerical vector is inputted to the conserved region detection process. Sections with high functional responses in the ancestor protein vector are essentially considered as conserved regions. Traditionally, a region is identified as the conserved region if the functional responses of the residues in the region are greater than a user specified threshold value. Although this approach is easy to implement, it tends to tear one conserved region into more than two separated regions if any residue response is below the threshold value. This makes the approach tends to generate many short and discontinue conserved regions since the functional responses of adjacent residues are ignored.

To avoid the drawback, this research applies Haar wavelet transform technique (Haar, 1910) to capture the different degrees of gross features of the common ancestor vector at different scales during the detection process. The Haar wavelet transform uses recursive averaging and differencing operations to act on the elements of the ancestor vector, so that the ancestor vector can be smoothed while keeping the detail information needed to reconstruct the original ancestor vector from the smoothed version. In the beginning, at the scale 0, the common ancestor vector is considered as the initial scaling coefficient vector $\boldsymbol{c}_0 = [c_0^i | i = 1, \cdots, n]$ where c_0^i is the ith element of $\boldsymbol{c}_0$ and n is the length of $\boldsymbol{c}_0$. Then, the scaling coefficient vector $\boldsymbol{c}_j = [c_j^i | i = 0, \cdots, \frac{n}{2^j} - 1]$, the smooth version of $\boldsymbol{c}_0$ at the scale j ($j = 1, \cdots, \log_2 n$), is recursively computed by averaging pairs of consecutive elements of $\boldsymbol{c}_{j-1}$. The averaging operation on pairs of consecutive elements of $\boldsymbol{c}_{j-1}$ to generate the elements of $\boldsymbol{c}_j$ is shown as Eq. (3). Similarly, the wavelet coefficient vector $\boldsymbol{d}_j = [d_j^i | i = 0, \cdots, \frac{n}{2^j} - 1]$, the detail version of $\boldsymbol{c}_0$ at the scale j ($j = 1, \cdots, \log_2 n$), is recursively computed by differencing pairs of consecutive elements of $\boldsymbol{d}_{j-1}$. The differencing operation on pairs of consecutive elements of $\boldsymbol{d}_{j-1}$ to generate the elements of

$\boldsymbol{d}_j$ is shown as Eq. (4). Note that the length of $\boldsymbol{c}_j$ and $\boldsymbol{d}_j$ equal to the half length of $\boldsymbol{c}_{j-1}$,

$$c_j^i = \frac{1}{2}(c_{j-1}^{2i} + c_{j-1}^{2i+1}) \tag{3}$$

$$d_j^i = \frac{1}{2}(d_{j-1}^{2i} - d_{j-1}^{2i+1}) \tag{4}$$

At each scale j, an approximation vector $\boldsymbol{A}_j$ can be generated by duplicating each element of $\boldsymbol{c}_j$ $(j+1)$ times after $\boldsymbol{c}_j$ has been obtained. $\boldsymbol{A}_j$ represents the vector that describes the smoothed result of the common ancestor vector $\boldsymbol{c}_0$ at the scale j. Note that the lengths of all approximation vectors at all scales are the same with the length of $\boldsymbol{c}_0$, n. Accordingly, the conserved region detection at different scales can be conducted based on all these approximation vectors. The transformation process is continued until the length of $\boldsymbol{c}_j$ equals to 1 or the total lengths of all conserved regions detected from $\boldsymbol{A}_j$ is less than the half length of $\boldsymbol{c}_0$. For example, let an ancestor protein vector $\boldsymbol{c}_0$ be $(9,7,3,5)$. At the scale 1, the scaling coefficient vector $\boldsymbol{c}_1$ is $(8,4)$ after taking the average of elements $\{9,7\}$ and $\{3,5\}$ in $\boldsymbol{c}_0$, respectively, whereas the wavelet coefficient vector $\boldsymbol{d}_1$ is $(1,-1)$ by taking the average difference of elements $\{9,7\}$ and $\{3,5\}$ in $\boldsymbol{c}_0$, respectively. Accordingly, the first approximation vector $\boldsymbol{A}_1$ of the ancestor protein vector is obtained as $(8,8,4,4)$ by duplicating each element in $\boldsymbol{c}_1 = (8,4)$ twice. Similarly, at the scale 2, the scaling coefficient vector $\boldsymbol{c}_2 = (6)$ is calculated by taking the average of elements $\{8, 4\}$ in $\boldsymbol{c}_1$, and the second approximation vector $\boldsymbol{A}_2 = (6,6,6,6)$ is generated by duplicating single element in $\boldsymbol{c}_2 = (6)$ four times. Because the length of $\boldsymbol{c}_2$ equals to 1, the transform process is finished. Fig. 6 illustrates the procedure of applying Haar wavelet transform to generate its two approximation vectors $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$.

Let μ_j be he average of all n elements in $\boldsymbol{A}_j$ at the scale j and be formulated by the following equation:

$$\mu_j = \frac{1}{n} \times \sum_{k=1}^{n} x_{jk} \tag{5}$$

where x_{jk} is the value of the kth element in $\boldsymbol{A}_j$. The average μ_j is then used to determine whether each element x_{jk} in $\boldsymbol{A}_j$ is included within the conserved regions at the jth scale or not. If $x_{jk} > \mu_j$, x_{jk} can be within the conserved regions at the jth scale; otherwise, x_{jk} can not within the region. The average-based criterion makes the conserved region detections at each scale be very efficient. Fig. 7 shows the procedure of using the average-based criterion to progressively detect conserved regions at different scales. At the scale 2 in Fig. 7, for example, there are four sections of residues in which their EIIP values are larger than the average 0.0204. They are from the 1st residue to the 12th residue, from the 17th residue to the 28th residue, from the 52th residue to the 56th residue, and from the 70th residue to the 80th residue. Therefore, these four sections are considered as conserved regions at the scale 2. In addition, the sum of lengths of these conserved regions detected at the scale 2 is 40 $(=12+12+5+11)$, less than the half length of the ancestor protein vector $(=88 \times 1/2)$, so that the progressive detection process is stopped at the scale 2.

Finally, all conserved regions detected at different scales are aggregated together to form final conserved regions for the ancestor protein vector. The aggregation is similar to the "OR" logic operation in which the result is TRUE (i.e. one) if the value of an element in any region is TRUE (i.e. one or non-zero). Note that the conserved regions detected from the first approximation vector $\boldsymbol{A}_1$ are not aggregated since the length of a conserved region at the scale 1 generally is less than three resides. For example, as shown in Fig. 7, the progressive detection process for the five example proteins is stopped at the scale 2. Therefore, no aggregation is required in this case. Fig. 8 shows another aggregation example using different protein dataset. It is found that there are four conserved regions at scale 2 and two conserved regions at scale 3. After aggregating the conserved regions at the two scales, three regions are detected as the common conserved regions of the new protein dataset.

3. Implementations

To show the performance of the proposed conserved region detection method, nine datasets in the first reference set of BAliBASE database are served as benchmark data since the conserved regions in each dataset have been identified using the PDBsum database (Laskowski et al., 1997), and aligned as well as the secondary structure elements by

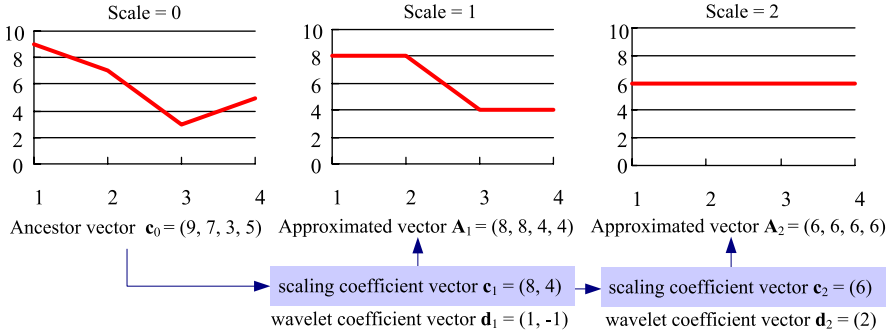

Fig. 6. Two approximation vectors generated by Haar wavelet transform.

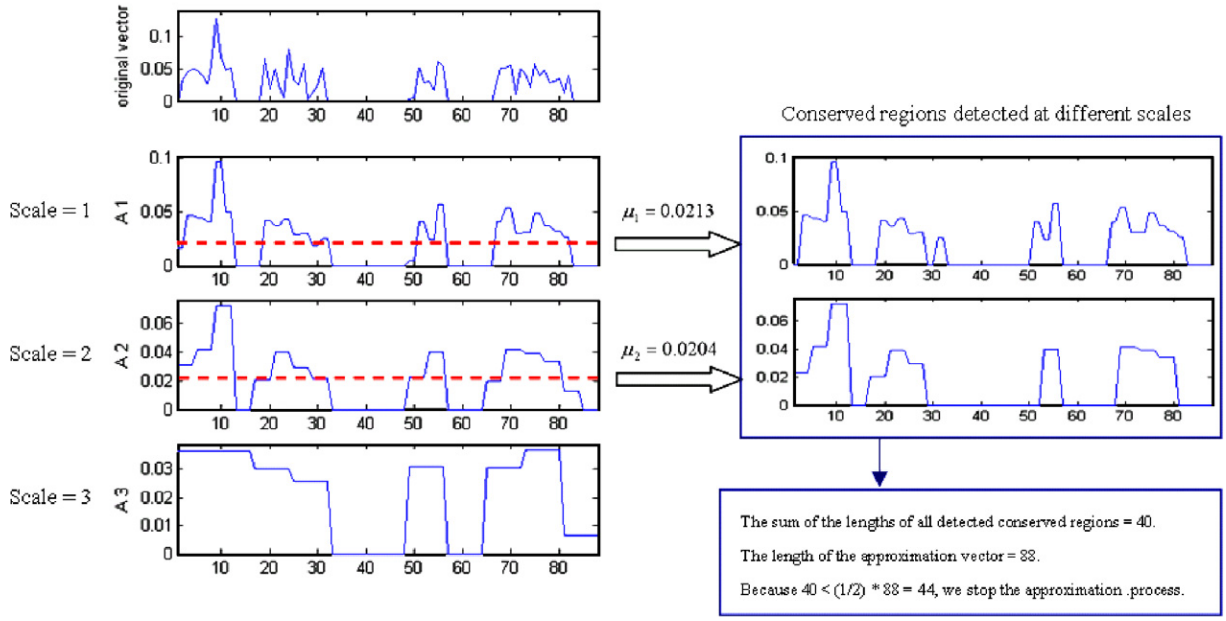

Fig. 7. Using the average-based criterion to progressively detect conserved regions.

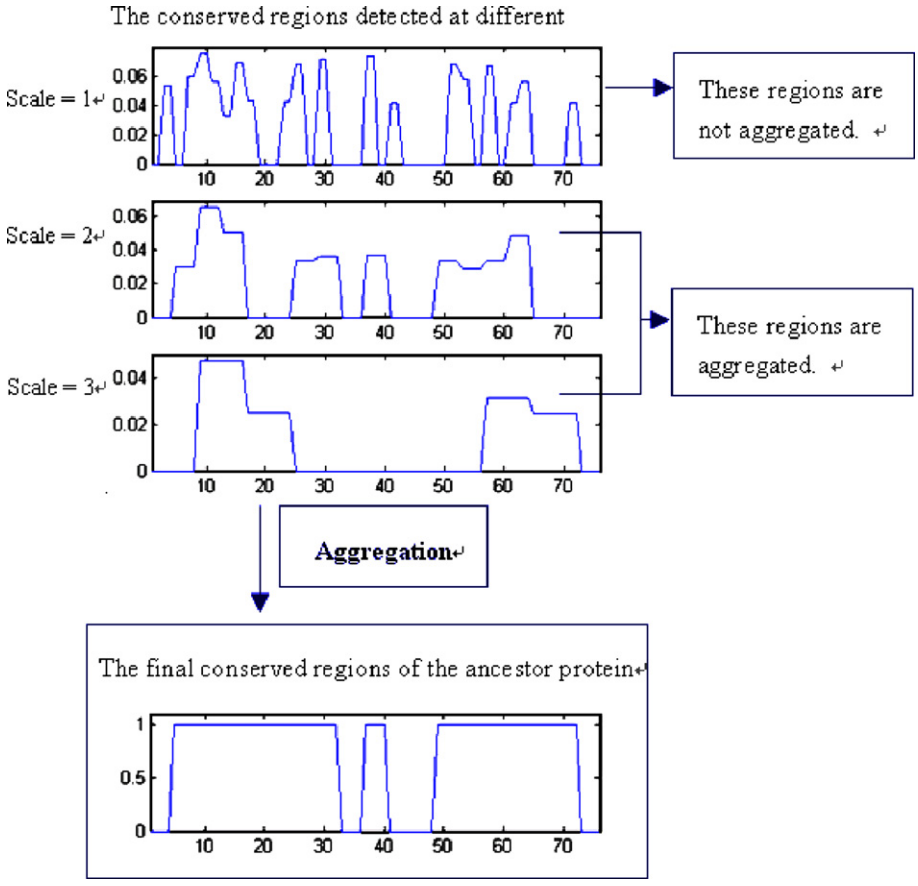

Fig. 8. The aggregation process of merging all conserved regions at different scales.

manual verification and adjustment. Table 1 describes the characteristics of the nine datasets. In addition, a popular conserved region detection tool, named Scorecons (http://www.ebi.ac.uk/thornton-srv/databases/cgi-bin/valdar/

Table 1
The characteristics of the nine datasets

Dataset name	Type	Number of sequences	Alignment length	Average percent identity (%)
1aboA (SH3)	Short, <25% identity	5	88	18
1aab (high mobility group protein)	Short, 20–40% identity	4	82	30
1aho (toxin II)	Short, >35% identity	5	76	44
1bbt3 (foot-and-mouth disease virus)	Medium, <25% identity	5	233	14
1ad2 (ribosomal protein l1)	Medium, 20–40% identity	4	221	31
1amk (triose phosphate isomerase)	Medium, >35% identity	5	258	49
1ajsA (aminotransferase)	Long, <25% identity	4	447	14
1ac5 (carboxypeptidase)	Long, 20–40% identity	4	524	29
1ad3 (aldehyde dehydrogenase)	Long, >35% identity	4	459	47

scorecons_server.pl), is compared in this study. The detail introduction of Scorecons tool can be referred in (Valdar, 2002).

Scorecons and our proposed methods both apply ClustalW to conduct the MSA task so that their alignment results are the same. All parameters in Scorecons are set as defaults suggested by the developer. If the conservation score of an aligned column given by Scorecons is larger than the average of conservation scores of all aligned columns, the column is included within the conserved region by Scorecons. Two common measures *Precision* and *Recall*, widely used in the domain of information retrieval (Witten and Frank, 1999), are employed to evaluate the performance for the two methods. Precision and recall measures are calculated based on a confusion matrix, which is illustrated in Fig. 9. Note that the true locations of conserved regions occurred on all protein sequences in each dataset have been reliably marked with the underscored columns, i.e. the core block in the alignment result for each dataset shown in the BAliBASE database. The higher the precision and recall measures, the higher the performance of a method is.

The precision and recall measures of the two methods after testing the nine datasets are listed, respectively, in Table 2. As shown in Table 2, the average precision measure of the proposed method is 79.24%, whereas the average precision measure of Scorecons is 59.75%. We further conduct the statistical paired *t*-test (Goulden, 1956) for the precision measures of the two methods and the test result shows the proposed method is significantly superior to Scorecons in terms of precision measure since the obtained *P*-value is less than 0.001. On the other hand, the recall measure of the proposed method is 60.45% which is smaller than 65.46% of Scorecons. However, when the two methods are compared in terms of the recall measure using the statistical paired *t*-test, the obtained *P*-value equals to 0.323, which means the proposed method is not significantly inferior to Scorecons in terms of recall measure because the *P*-value is larger than 0.05. Overall, the proposed method performs better in detecting conserved regions than Scorecons.

To know the influences of datasets with different sequence lengths, the nine datasets are grouped into three types including short, medium and long sequence lengths. The type of each dataset based on sequence length can be found in Table 1. For example, the datasets of 1aboA, 1aab and 1aho are the type of short sequence length. Figs. 10 and 11 illustrate the precision and recall measures, respectively, for the three protein sequence lengths using Scorecons and the proposed method. As illustrated in Fig. 10, the precision measures of Scorecons and the proposed method both tend to be downward slightly as the protein sequence length increasing. With the statistical paired *t*-test for the proposed method, it shows no significant influences of different protein sequence lengths on

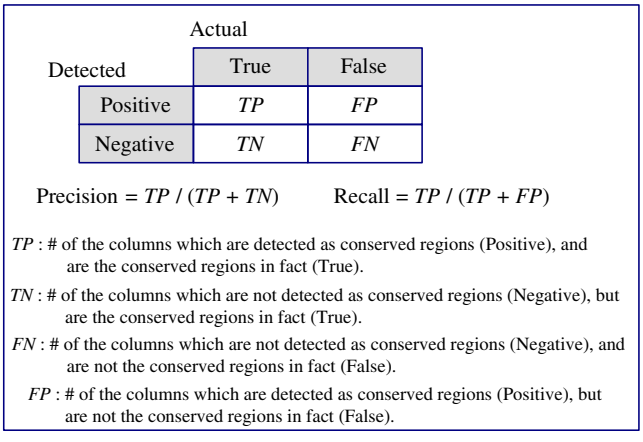

Fig. 9. The definition of precision and recall measures.

Table 2
The precision and recall measures for the nine datasets using two conserved region detection methods

Dataset name	Scorecons		The proposed method	
	Precision (%)	Recall (%)	Precision (%)	Recall (%)
1aboA	60.61	41.67	81.82	67.50
1aab	57.14	70.59	76.34	65.38
1aho	65.71	76.67	91.43	64.14
1bbt3	71.70	32.48	90.57	30.57
1ad2	55.15	75.76	71.62	52.86
1amk	54.22	98.39	79.56	94.53
1ajsA	45.45	32.43	60.15	35.58
1ac5	70.68	67.28	82.74	54.17
1ad3	57.10	93.83	78.91	79.34
Average	59.75	65.46	79.24	60.45

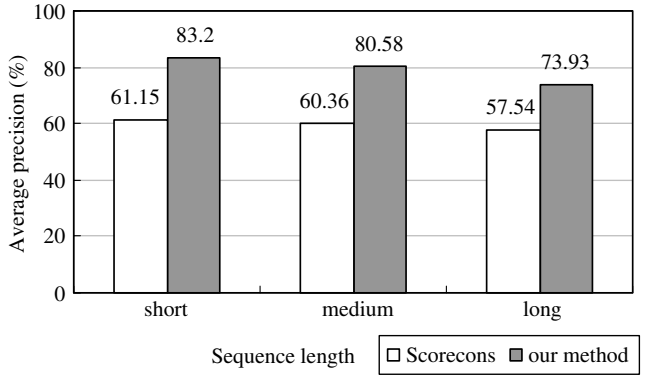

Fig. 10. Average precision comparisons for different protein sequence lengths.

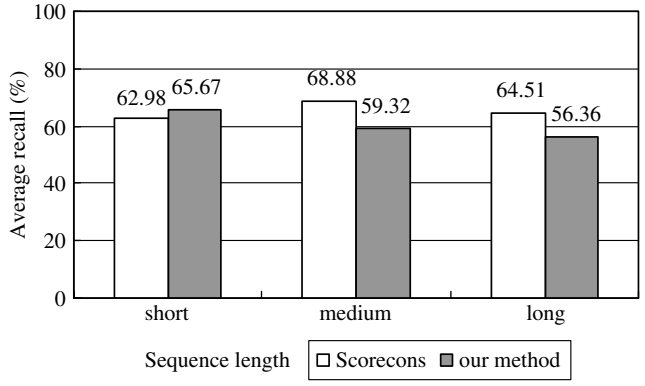

Fig. 11. Average recall comparisons for different protein sequence lengths.

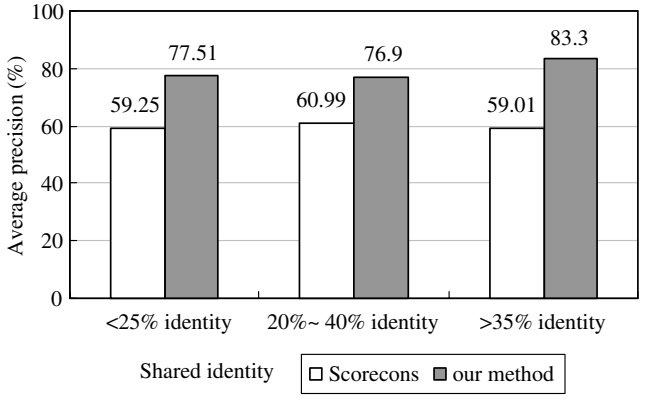

Fig. 12. Average precision comparisons for different degrees of shared identities.

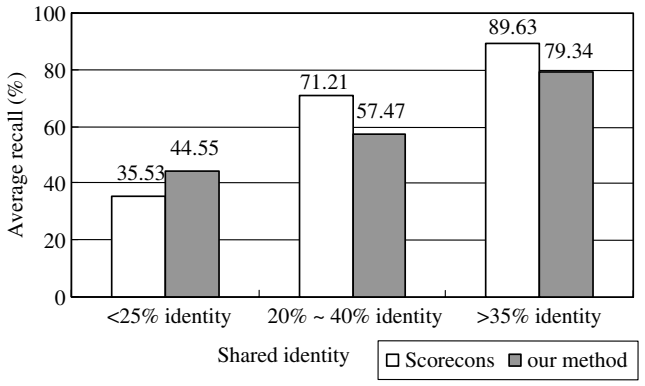

Fig. 13. Average recall comparisons for different degrees of shared identities.

the precision measure. When the protein sequence length increases, the recall measure of the proposed method declines gradually but the one of Scorecons fluctuates unsteadily, depicted as Fig. 11. Similarly, with the statistical paired *t*-test for the proposed method, it shows no significant influences of different protein sequence lengths on the recall measure. Even if the length of a protein sequence does not significantly influence on the two measures of the proposed method, these experiments show the two measures still decreases slightly as the length of a protein sequence increases.

Except the influence of sequence length, degree of shared identity might result in critical impact to the performance of the proposed method. To know the influence, the nine datasets are partitioned into three identity types, including "<25% identity", "20–40% identity" and ">35% identity". The grouping result for each dataset based on degree of shared identity can be found in Table 1. For instance, 1aobA, 1bbt3 and 1ajsA are the type of "<25% identity". Figs. 12 and 13 summarize the precisions and recalls for the three degrees of shared identities using Scorecons and the proposed method, respectively. As depicted in Fig. 12, when the degree of shared identity increases, the precision measure of the proposed method increases slightly, but the precision measure of Scorecons remains constant. On the other hand, in Fig. 13 the recall measure of the two methods both increase gradually when the degree of shared identity increases. With the statistical paired *t*-test for the proposed method, it shows no significant influences of different the degree of shared identity on the precision measure, but there is a significant influence on the recall measure. That is, the higher the degree of shared identity, the higher the recall measure for detecting conservation regions is. In addition, in the situation of low degree of shared identity (i.e. "<25% identity"), the precision and recall measures of the proposed method (i.e. 77.51% and 44.55%) are obviously greater than the ones of Scorecons. It means that the proposed method has high reliability to deal with the residue divergence problem (Rost, 1999).

Finally, we examine the performances of the proposed method using different phylogenetic tree construction approaches since they maybe result in different common ancestor protein vectors. Besides the neighbor-joining approach adopted in ClustalW, the maximum parsimony (MP) (Felsenstein, 1978) and the maximum likelihood (ML) (Felsenstein, 1981) approaches are also frequently used to construct the unrooted phylogenetic tree. PHYLIP (http://evolution.gs.washington.edu/phylip.html) is a

Table 3
The precision and recall measures for the nine datasets using three phylogenetic tree construction approaches

Dataset name	MP		ML		Neighbor-joining	
	Precision (%)	Recall (%)	Precision (%)	Recall (%)	Precision (%)	Recall (%)
1aboA	79.64	64.32	80.31	66.45	81.82	67.50
1aab	74.72	63.86	74.72	63.86	76.34	65.38
1aho	91.43	64.14	91.43	64.14	91.43	64.14
1bbt3	***86.27***	***28.85***	***87.34***	***29.16***	***90.57***	***30.57***
1ad2	70.82	51.48	70.82	51.48	71.62	52.86
1amk	79.56	94.53	79.56	94.53	79.56	94.53
1ajsA	***56.78***	***33.62***	***57.18***	***32.57***	***60.15***	***35.58***
1ac5	***78.52***	***52.75***	***78.52***	***52.75***	***82.74***	***54.17***
1ad3	77.65	77.23	77.65	77.23	78.91	79.34
Average	77.27	58.98	77.50	59.13	79.24	60.45

package of programs for inferring an unrooted phylogenetic tree which can apply these three approaches, respectively. After executing each of the three approaches, the root of the unrooted phylogenetic trees can be placed using the mid-point manner (Thompson et al., 1994b). In PHYLIP, the inputted data for MP and ML approaches require the alignment result for multiple protein sequences, which is different from the pairwise similarity scores using in the neighbor-joining approach. For fair comparison in this experiment, therefore, ClustalW is applied to generate not only the sequence alignment result for MP and ML approaches but also the pairwise similarity scores for neighbor-joining approach. The precision and recall measures of the three phylogenetic tree construction approaches after testing the nine datasets are listed, respectively, in Table 3.

As shown in Table 3, when protein sequences have long lengths or low degree of shares identity, such as 1bbt3, 1ajsA and 1ac5 datasets, the precision and recall measures of the neighbor-joining approach are superior to the results of MP and ML approaches. In other datasets, the performances using the three approaches are similar or equal. For this research, therefore, the neighbor-joining approach is an appropriate choice to construct the phylogenetic tree.

4. Demonstration for a practical case

After evidencing the performance of the proposed conserved region detection method, the method is subsequently implemented for a real case to show its usage feasibility. The Kabat database (Kabat et al., 1991) aims at identifying the antibody combining site based on available protein sequences. In the Kabat database, nine protein sequences categorized in the human immunoglobulin λ-chain V-I region, including l1hung, l1hubl, l1hunw, l1huep, l1huwa, l1humm, l1huha, l1huvo, and l1hunm, are selected as the practical case in our demonstration.

First, the nine protein sequences are aligned using the ClustalW tool, and the result of its multiple sequence alignment and phylogenetic tree are shown as Fig. 14. Then, the nine aligned protein sequences are transformed into nine numerical vectors according to the EIIP value of each amino acid, which is shown as Fig. 15. Subsequently, the EIIP vector of the common ancestor protein is inferred using a serial of PCA approaches by following the branching order in the phylogenetic tree. The extracted EIIP vector of the common ancestor protein is shown as Fig. 16.

After extracting the common ancestor protein vector, the Haar wavelet transform technique is used to detect the conserved regions at different scales, including seven conserved regions at scale 2 and three conserved regions at scale 3. After aggregating the conserved regions at the two scales, four regions are detected as the common conserved regions of the nine protein sequences. Fig. 17 shows the detected conserved regions at scale 2 and scale 3 and the common conserved regions through the aggregation step.

Finally, we compare the locations of conserved regions detected by the proposed method, i.e. the four subsequences surrounded by red boxes in Fig. 18, and the true conserved regions annotated in the Kabat database, i.e.

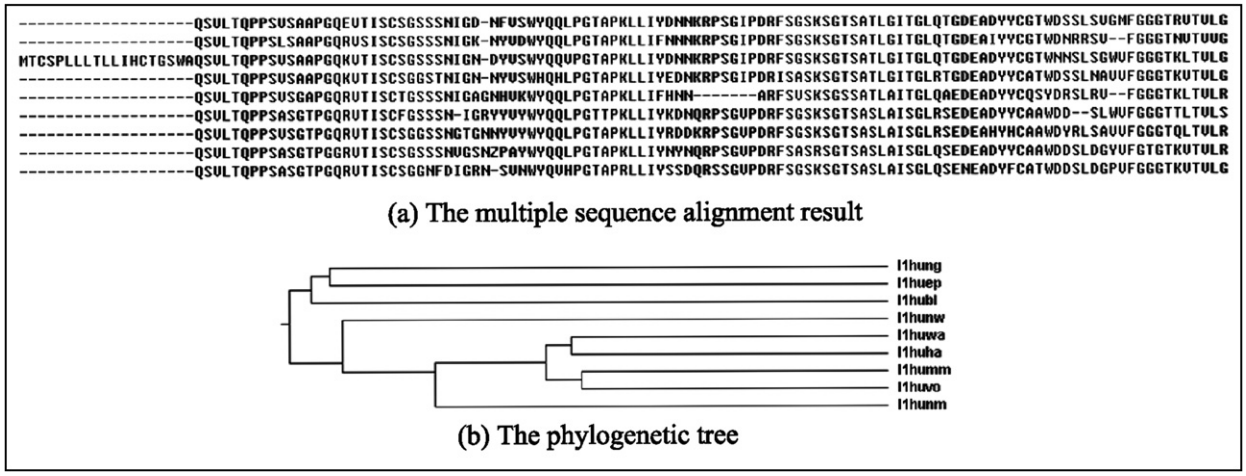

Fig. 14. Multiple sequence alignment for the nine protein sequences using ClustalW.

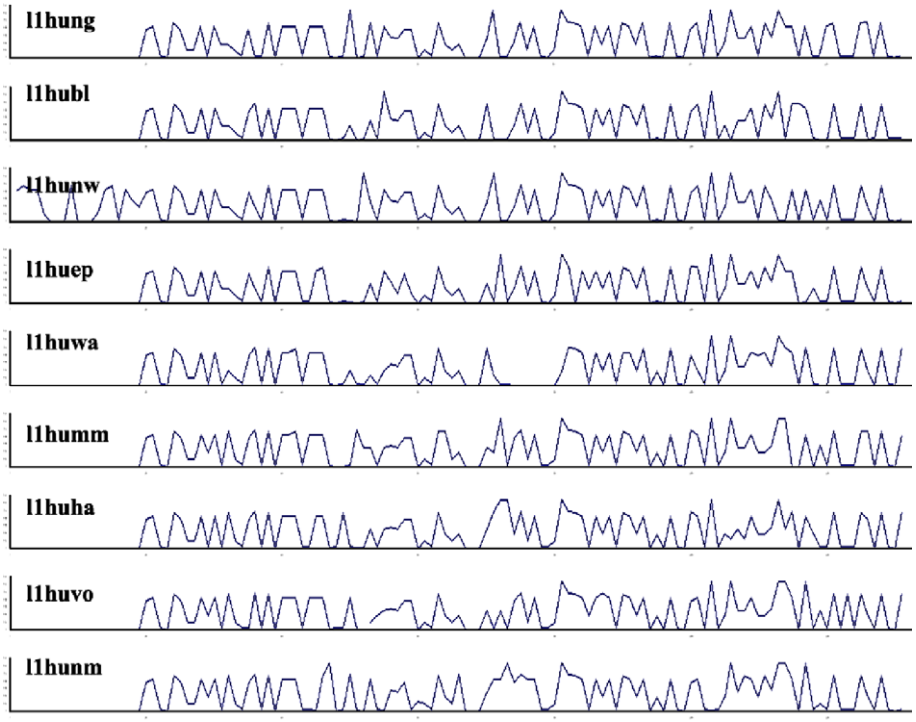

Fig. 15. Transforming the nine aligned protein sequences into nine numerical vectors.

Fig. 16. The EIIP vector of the common ancestor protein extracted through a serial of PCA.

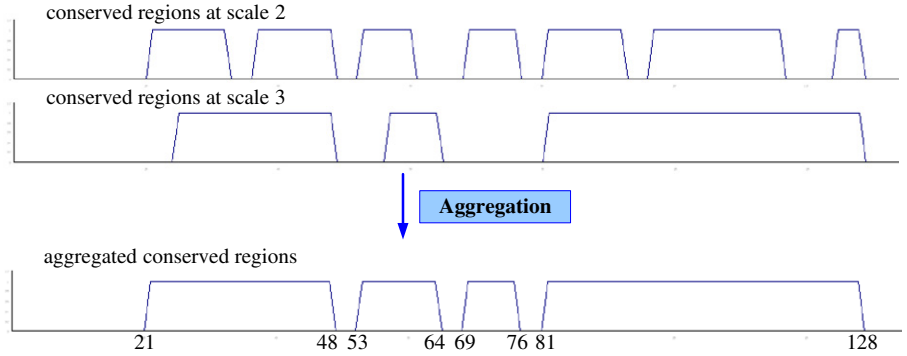

Fig. 17. The conserved regions detected at scale 2 and scale 3 and their aggregation result as the common conserved regions.

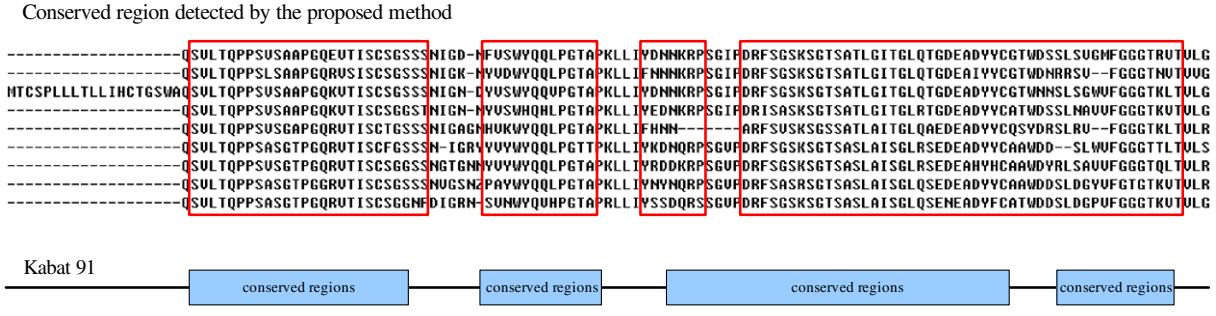

Fig. 18. The comparison between the locations of conserved regions detected by the proposed method and the ones annotated in the Kabat database (Kabat et al., 1991).

the locations of the four blue boxes on the Kabat91 sequence line in Fig. 18. As illustrated in Fig. 18, the first two and last true conserved regions are correctly detected by the proposed method, but only a short section in the third true conserved regions can not be fully recognized. In addition, only a short false-detecting section occurred in the last conserved region through the proposed method. The calculated Precision and Recall measures are 95.6% and 81.8%, respectively. In short, even though there are two small errors in the detection result of the proposed method in this case, the usage feasibility of the proposed method is still reliable enough to be used in practice. Especially, the usage of the proposed method does not require artificial involvements for setting parameters, so that it is convenient for users to manipulate the proposed method.

5. Conclusions

A conserved region composed of a number of successive amino acids carries a specific functional importance for a protein sequence. Therefore, conserved region detection from multiple protein sequences has been one of the major focuses in the fields of bioinformatics. Instead of deriving conserved regions by considering all original protein sequences simultaneously, this research focus the conserved region detection on the common ancestor protein of these protein sequences. At the beginning of the proposed method, all protein sequences are aligned using a specific multiple sequence alignment (MSA) tool and transformed into numerical vectors in advance. In the process of ancestor inference, principal component analysis (PCA) approach is used to infer the hypothesized proteins from the phylogenetic tree generated during MSA. Then, Haar wavelet transform is employed to detect the conserved regions of the inferred ancestor protein vector at different scales. Consequently, the detected conserved regions are considered as the common conserved regions of the original protein sequences. A set of experiments indicate the proposed method is more precise and efficient than a popular detection tool – Scorecons in conserved region detection. In addition, it is confirmed that the length of sequence and the degree of shared identity both influence on the performance of the proposed method.

The performance of the proposed method depends on the inference quality of the common ancestor protein vector which is inferred based on a hypothetical evolutionary history. Different evolution process will generate a different common ancestor protein vector. Through our experiments, the neighbor-jointing approach is more proper than the MP and ML approaches for this research. Furthermore, the usage feasibility of the proposed method is reliable enough to be used in practice through our demonstration for a practical case.

In the future, we will further study how to construct a phylogenetic tree which describes the most meaningful evolutionary history of the participatory proteins. In addition, for a protein family, the conserved regions detected based on various functional responses can be considered as the family features. With these family features, protein analysis tasks such as unknown protein classification, protein structure prediction, and protein family annotation can be conducted more efficiently and effectively. Therefore, using various functional responses to generate a number of family features will be another interesting research topic in the future.

References

Armon, A., Graur, D., Ben-Tal, N., 2001. ConSurf: An algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. J. Mol. Biol. 307 (1), 447–463.

Bashford, D., Chothia, C., Lesk, A.M., 1987. Determinants of a protein fold: Unique features of the globin amino acid sequences. J. Mol. Biol. 196 (1), 199–216.

Chan, S.C., Wong, A.K., Chiu, D.K., 1992. A survey of multiple sequence comparison methods. Bull. Math Biol. 54 (4), 563–598.

Cosic, I., 1994. Macromolecular bioactivity: Is it resonant interaction between macromolecules? – Theory and applications. IEEE Trans. Biomed. Eng. 41 (12), 1101–1114.

Felsenstein, J., 1978. Cases in which parsimony and compatibility methods will be positively misleading. Syst. Zool. 27 (4), 401–410.

Felsenstein, J., 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach. J. Mol. Evol. 17 (6), 368–376.

Fortuna, J., Capson, D., 2004. Improved support vector classification using PCA and ICA feature space modification. Pattern Recognition 37 (6), 1117–1129.

Gerstein, M., Altman, R.B., 1995. Average core structures and variability measures for protein families: Application to the immunoglobulins. J. Mol. Biol. 251 (1), 161–175.

Goulden, C.H., 1956. Methods of Statistical Analysis. Wiley, New York, 50–55.

Guru, D.S., Punitha, P., 2004. An invariant scheme for exact match retrieval of symbolic images based upon principal component analysis. Pattern Recognition Lett. 25 (1), 73–86.

Haar, A., 1910. Zur theorie der orthogonalen funktionensysteme. Math. Ann. 69, 331–371.

Jolliffe, I.T., 1986. Principal Component Analysis. Springer-Verlag, New York.

Jores, R., Alzari, P.M., Meo, T., 1990. Resolution of hypervariable regions in T-cell receptor chains by a modified Wu-Kabat Index of amino acid diversity. Proc. Natl. Acad. Sci. USA 87 (23), 9138–9142.

Kabat, E.A., Wu, T.T., Perry, H., Gottesman, K., Foeller, C., 1991. Sequences of Proteins of Immunological Interest, fourth ed. NIH Publication No. 91-3242.

Karlin, S., Brocchieri, L., 1996. Evolutionary conservation of RecA genes in relation to protein structure and function. J. Bacteriol. 178 (7), 1881–1894.

Landgraf, R., Fischer, D., Eisenberg, D., 1999. Analysis of Heregulin symmetry by weighted evolutionary tracing. Protein Eng. 12 (11), 943–951.

Laskowski, R.A., Hutchinson, E.G., Michie, A.D., Wallace, A.C., Jones, M.L., Thornton, J.M., 1997. PDBsum: A web-based database of summaries and analyses of all PDB structures. Trends Biochem. Sci. 22 (12), 488–490.

Lichtarge, O., Bourne, H.R., Cohen, F.E., 1996. An evolutionary trace method defines binding surfaces common to protein families. J. Mol. Biol. 257 (2), 342–358.

Lockless, S.W., Ranganathan, R., 1999. Evolutionarily conserved pathways of energetic connectivity in protein families. Science 286, 295–299.

Mirny, L.A., Shakhnovich, E.I., 1999. Universally conserved positions in protein folds: Reading evolutionary signals about stability, folding kinetics and function. J. Mol. Biol. 291 (1), 177–196.

Orengo, C.A., Jones, D.T., Thornton, J.M., 2003. Bioinformatics: Genes, Proteins and Computers. BIOS Scientific Publishers Ltd., Oxford, UK.

Rost, B., 1999. Twilight zone of protein sequence alignments. Protein Eng. 12 (2), 85–94.

Saitou, N., Nei, M., 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4 (4), 406–425.

Sander, C., Schneider, R., 1991. Database of homology-derived protein structures and the structural meaning of sequence alignment. Proteins 9 (1), 56–68.

Shannon, C.E., 1948. A mathematical theory of communication. Bell System Tech. J. 27, 379–423, 623–656.

Shenkin, P.S., Erman, B., Mastrandrea, L.D., 1991. Information-theoretical entropy as a measure of sequence variability. Proteins 11 (4), 297–313.

Thompson, J.D., Higgins, D.G., Gibson, T.J., 1994a. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22 (22), 4673–4680.

Thompson, J.D., Higgins, D.G., Gibson, T.J., 1994b. Improved sensitivity of profile searches through the use of sequence weights and gap excision. Bioinformatics 10 (1), 19–29.

Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F., Higgins, D.G., 1997. The CLUSTALX windows interface: Flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. 25 (24), 4876–4882.

Thompson, J.D., Plewniak, F., Poch, O., 1999. BAliBASE: A benchmark alignment database for the evaluation of multiple alignment programs. Bioinformatics 15 (1), 87–88.

Valdar, W.S., 2002. Scoring residue conservation. Proteins 48 (2), 227–241.

Williamson, R.M., 1995. Information theory analysis of the relationship between primary sequence structure and ligand recognition among a class of facilitated transporters. J. Theor. Biol. 174 (2), 179–188.

Witten, I.H., Frank, E., 1999. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann, San Francisco.

Wu, T.T., Kabat, E.A., 1970. An analysis of the sequences of the variable regions of Bence Jones proteins and myeloma light chains and their implications for antibody complementarity. J. Exp. Med. 132, 211–250.