Available online at www.sciencedirect.com
ScienceDirect
Pattern Recognition 40 (2007) 516–529
PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY
www.elsevier.com/locate/patcog

Spectral distortion measures for biological sequence comparisons and database searching

Tuan D. Pham[a, b, *]

[a] *Bioinformatics Applications Research Center, James Cook University, Townsville, QLD 4811, Australia*
[b] *School of Information Technology, James Cook University, Townsville, QLD 4811, Australia*

Received 21 July 2005; received in revised form 19 February 2006; accepted 28 February 2006

Abstract

In bioinformatics and computational biology, methods for biological sequence comparison play the most important role for the interpretation of complex nucleotide and protein data such as the inference of relationships between genes, proteins and species; and the discovery of novel protein structures and functions. This type of inference is derived by sequence similarity matching on the databases of biological sequences. As many entire genomes have being determined at a rapid rate, computational methods for comparing genomic and protein sequences will be more essential for probing the complexity of genes, genomes, and molecular machines. In this paper we introduce a pattern-comparison algorithm, which is based on the mathematical concepts of linear predictive coding and its cepstral-distortion measures for the analyses of both DNA and protein sequences. The results obtained from several experiments on real datasets have shown the effectiveness of the proposed approach.

Keywords: Distortion measures; DNA sequences; Protein sequences; Bioinformatics

1. Introduction

Finding the similarities between related DNA or protein sequences helps life-science researchers understand the information content and functions of biological sequences. For example, by comparing DNA sequences based on their similarity measures, the function of a new genetic sequence or the evolutionary relationships among a set of similar genes can be determined. Genes have attracted significant attention from the community of computational biology and bioinformatics; it is the proteins which perform essential roles for controlling, effecting and modulating biochemical, cellular, and phenotypic functions. A novel protein function can be inferred from the known functions of homologous proteins and this type of prediction is based on the similar sequence-similar structure-similar function paradigm [1]. Protein sequence comparison is the most powerful tool for such inference and analysis [2].

Such above understanding of biological information will lead to the discovery of the pathways involved in normal processes and in disease pathogenesis. As a result, many disorders can be correctly diagnosed, treated and prevented. Therefore, the development of effective methods for comparing protein sequences is one of the major tasks of biological research for gaining better understanding of the complexity of molecular machines.

Given the importance of research into methodologies for computing similarity among biological sequences, there have been a number of computational and statistical methods for sequence comparisons developed over the past decade. However, it still remains a challenging problem for the research community of computational biology [3,4], because there is no single method that can provide reliable solutions to a variety of problems. Two distinct bioinformatic methodologies for studying the similarity/dissimilarity of sequences are known as alignment-based and alignment-free methods.

* Corresponding author at: School of Information Technology, James Cook University, Townsville, QLD 4811, Australia. Tel.: +61 7 47816903.
E-mail address: tuan.pham@jcu.edu.au.

0031-3203/$30.00 © 2006 Pattern Recognition Society. Published by Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2006.02.026

Methods for comparing of sequences by alignment have been utilized for solving many important problems in biology. The alignment of nucleotide (DNA) or amino acid (protein) sequences is aimed to discover the evolutionary relationship between homologs which are sequences sharing a common ancestor. A simple alignment between a pair of sequences can be performed by matching the characters of the two sequences. While matching the characters, alignment methods take into account three kinds of changes that can occur at any position within a sequence [5]: mutation (replacing one character with another), insertion (adding one or more positions), and deletion (deleting one or more positions). Gaps can usually be added in the alignments to express insertion or deletion. Since there are many possible ways of aligning sequences, a common optimal alignment is based on the maximization of a scoring function which assigns a credit (positive score) to each aligned pair of identical residues called the match score, and a penalty (negative score) to each aligned pair of nonidentical residues called the mismatch score. The assignment of match and mismatch scores for ungapped positions in sequence alignment is based on one of several derived scoring matrices such as the identity matrix, BLAST matrix, and transition–transversion matrix, whose values are heuristically determined based on the genetic codes, the observed chemical or physical similarity, and observed substitution rates among residues [5].

Due to the indel (insertion/deletion) events, there are many possible alignments between two or more sequences which make the finding of the best alignment using exhaustive search unfeasible. Dynamic programming [6] is the first optimization approach being applied for solving this searching problem of sequence alignments: the Needleman and Wunsch algorithm [7] (global alignment), and the Smith–Waterman algorithm [8] (local alignment). There are many other optimization methods recently developed for sequence alignments such as those based on hidden Markov models [9].

Regarding to the application of alignment methods for the searching for database-stored nucleotide or protein sequences that are similar to the nucleotide or protein query sequence respectively, BLAST (Basic Local Alignment Search Tool) [10] and its derivatives are among the most well known tools for such task, which performs the searching based on ungapped local alignments. In order to make the searching efficient, a strategy is to break the query sequence into words of a fixed length and BLAST searches for word matches in the database and extends the match until the local alignment score falls below a given threshold. For local alignments that allows gaps, FASTA [11] and its derivatives are popular tools for database searching. For multiple sequence alignment, CLUSTAL [12] and its variant CLUSTALW [13] are well known algorithms for such analysis; whereas the latter algorithm can improve the sensitivity of the progressive alignment by weighting, and positions-specific gap penalties.

The book by Krane and Raymer [5], which we have already cited above, provides an excellent and concise overview of sequence alignment methods for readers who are new to the field of bioinformatics.

Given the usefulness of alignment-based tools, it has been known that each conventional sequence comparison method has its own advantage and disadvantage. For alignment-based similarity measures, the term *alignment* can be changed to *edit*, and similarly, *alignment score* be changed to *edit distance*. Moreover, some edits can change the original function of the sequence. Hence, when there is an edit, some penalty should be given. The problem is that, in general, the lengths of biological sequences are quite different. For example, the length of human α hemoglobin (HAHU) and human basic FGF are 143 and 288, respectively. When the global pairwise alignment of these sequences are executed by using EMBOSS (http://www.ebi.ac.uk/emboss/align/), the gaps are about 70%. Therefore, it is clear that alignment-based similarity measures have some disadvantages. As a result, bioinformaticians have sought to find new methodologies for measuring similarity between biological sequences. One of the promising methodologies for dealing with biological data is the signal-processing based approach [14–18]. It has been pointed out by Anatassiou [19] that if protein or DNA sequences can be mapped into one or more numerical sequences, then digital signal processing would be very useful for solving highly relevant problems in bioinformatics and computational biology.

In addition, the search for optimal solutions using sequence alignment-based methods is encountered with difficulty in computational aspect with regard to large biological databases and long sequences. Therefore, the emergence of research into alignment-free sequence analysis is apparent and necessary to overcome critical limitations of sequence analysis by alignment.

Methods for alignment-free sequence comparison of biological sequences utilize several concepts of distance measures [20], such as the Euclidean distance [21], Euclidean and Mahalanobis distances [22], Markov chain models and Kullback–Leibler discrepancy (KLD) [23], cosine distance [24], Kolmogorov complexity [25], and chaos theory [26]. Our previous work [27] on sequence comparison has some strong similarity to the work by Wu et al. [23], in which statistical measures of DNA sequence dissimilarity are performed using the Mahalanobis distance and the standardized Euclidean distance under Markov chain model of base composition, as well as the extended KLD. The KLD extended by Wu et al. [23] was computed in terms of two vectors of relative frequencies of n-words over a sliding window from two given DNA sequences. Whereas, our previous work derives a probabilistic distance between two sequences using a symmetrized version of the KLD, which directly compares two Markov models built for the two corresponding biological sequences.

Cosic [28] developed the resonant recognition model (RRM), which is based on the mapping of a protein sequence into a numerical sequence in which each amino acid can be assigned with a real constant called the electron–ion

interaction potential (EIIP) value [29,30] to analyze protein interaction using the discrete Fourier transform. Trad et al. [31] applied the wavelet transform [32,33] for comparing protein sequences at different scales of the protein signals. Unlike the popular Fourier transform which maps the input data into a new space using the basis functions of sines and cosines, the wavelet transform maps the input signal into a new space using the basis functions which are localized in space. Thus the term *wavelet* is used to indicate a localized wave-like function. Wavelets are localized in frequency as well as in space; whereas the Fourier transform is not local in space but in frequency. As another remark, Fourier analysis is unique but wavelet analysis has many possible sets of wavelets. Wavelet features have been used for solving many pattern recognition problems [34] such as chromosome classification [35], image indexing [36], speech enhancement [37], and invariant pattern recognition [38].

The wavelet-based method studied by Trad et al. [31] transforms each protein sequence into a corresponding numerical sequence using the RRM. This numerical sequence is then normalized to have zero mean and unit deviation. Shorter sequences were zero-padded to have the same length of the longest sequence in order to enable the calculation of the cross-correlation coefficients. However, the zero-padding may cause error in defining peak frequencies, which leads to undesirable effect on the analysis of biological sequences [14]. Moreover, interpretation of similarities between two protein sequences in terms of multi-scales indicating contrasting results such as strong, weak, and no correlations which would be difficult for making decision.

In this paper we are interested in the novel application of some spectral distortion measures for nucleotide and protein sequence comparisons, where the computation is efficient and does not depend on sequence alignment. In the following sections we will firstly discuss how a sequence of nucleotides or amino acids can be represented as a sequence of corresponding numerical values; secondly we will then address how we can extract the spectral feature of these biological sequences using the method of linear predictive coding (LPC); thirdly we will present the concept of distortion measures of any pair of the sequences, which serve as the basis for the computation of sequence similarity. For DNA sequences, we have tested our method with six DNA sequences taken from *Escherichia coli* K-12 and *Shigella flexneri*, and one simulated sequence to discover their relations; and a complex set of 40 DNA sequences to search for most similar sequences to a particular query sequence. We have found that the results obtained from our proposed method are better than those obtained from other distance measures [23,27]. For protein sequences, the experimental results of this study have shown that LPC-based cepstral distortion measures could identify more related protein sequences than other existing methods such as alignment-based and wavelet-transform-based methods. An additional advantage of the new method is that it is physically reasonable and computationally tractable for DNA/protein sequence comparisons.

2. Representing biological sequences with EIIP values

In this section we will describe the EIIP values for nucleotides and amino acids. These numeral representations make it explicit for digital signal processing of biomolecular data using LPC; and LPC-based cepstral-distortion measures for sequence comparison.

The RRM is a physical and mathematical model which can extract protein or DNA sequences using signal analysis methods [29,28]. This approach can be divided into two parts. The first part involves the transformation of a biological sequence into a numerical sequence—each amino acid or nucleotide can be represented by the value of the EIIP [14] which describes the average energy states of all valence electrons in a particular amino acid or nucleotide. The EIIP values for each nucleotide or amino acid were calculated using the following general model pseudopotential [29,14,30,39]

$$\langle k + q[w]k\rangle = \frac{0.25Z\sin(\pi \times 1.04Z)}{2\pi}, \tag{1}$$

where q is a change of momentum of the delocalized electron in the interaction with potential w, and

$$Z = \frac{(\sigma Z_i)}{N}, \tag{2}$$

where Z_i is the number of valence electrons of the ith component, N is the total number of atoms in the amino acid or nucleotide. Each amino acid or nucleotide can be converted as a unique number, regardless of its position in a sequence. Thus, using the pseudopotential model, each nucleotide or amino acid located at any sequential position can be represented by a real EIIP constant which is the average energy states of all valence electrons in a particular amino acid. Table 1 shows the EIIP values for 20 amino acids.

3. Spectral features of biological sequences

It has been pointed out that the difficulty for the application of signal processing to the analysis of biological data is because it deals with numerical sequences rather than character strings [15,19]. If a character string can be converted into a numerical sequence, then digital signal processing can provide a set of novel and useful tools for solving highly relevant problems. By making use of the EIIP values for DNA or protein sequences, we will apply the principle of LPC to extract the spectral feature of the respective sequence known as the LPC cepstral coefficients.

3.1. Linear prediction coefficients

The estimated value of a particular nucleotide s_m at position or time n, denoted as $\hat{s}(n)$, can be calculated as a linear

Table 1
Electron–Ion Interaction Potential (EIIP) values for nucleotides and amino acids [28,31]

Nucleotide	EIIP
A	0.1260
G	0.0806
T	0.1335
C	0.1340
Amino acid	**EIIP**
Leu	0.0000
Ile	0.0000
Asn	0.0036
Gly	0.0050
Val	0.0057
Glu	0.0058
Pro	0.0198
His	0.0242
Lys	0.0371
Ala	0.0373
Tyr	0.0516
Trp	0.0548
Gln	0.0761
Met	0.0823
Ser	0.0829
Cys	0.0829
Thr	0.0941
Phe	0.0946
Arg	0.0959
Asp	0.1263

combination of the past p microarray samples. This linear prediction can be expressed as [40,41]

$$\hat{s}(n) = \sum_{k=1}^{p} a_k s(n-k), \tag{3}$$

where the terms $\{a_k\}$ are called the linear prediction coefficients (LPC).

The prediction error $e(n)$ between the observed sample $s(n)$ and the predicted value $\hat{s}(n)$ can be defined as

$$e(n) = s(n) - \hat{s}(n) = s(n) - \sum_{k=1}^{p} a_k s(n-k). \tag{4}$$

The prediction coefficients $\{a_k\}$ can be optimally determined by minimizing the sum of squared errors

$$E = \sum_{n=1}^{N} e^2(n) = \sum_{n=1}^{N} \left[s(n) - \sum_{k=1}^{p} a_k s(n-k) \right]^2. \tag{5}$$

To solve (5) for the prediction coefficients, we differentiate E with respect to each a_k and equate the result to zero

$$\frac{\partial E}{\partial a_k} = 0, \quad k = 1, \ldots, p. \tag{6}$$

The result is a set of p linear equations

$$\sum_{k=1}^{p} a_k r(|m-k|) = r(m), \quad m = 1, \ldots, p, \tag{7}$$

where $r(m-k)$ is the autocorrelation function of $s(n)$, that is symmetric, i.e. $r(-k) = r(k)$, and expressed as

$$r(m) = \sum_{n=1}^{N-m} s(n)s(n+m), \quad m = 0, \ldots, p. \tag{8}$$

Eq. (7) can be expressed in matrix form as

$$\mathbf{Ra} = \mathbf{r}, \tag{9}$$

where $\mathbf{R}$ is a $p \times p$ autocorrelation matrix, $\mathbf{r}$ is a $p \times 1$ autocorrelation vector, and $\mathbf{a}$ is a $p \times 1$ vector of prediction coefficients

$$\mathbf{R} = \begin{bmatrix} r(0) & r(1) & r(2) & \cdots & r(p-1) \\ r(1) & r(0) & r(1) & \cdots & r(p-2) \\ r(2) & r(1) & r(0) & \cdots & r(p-3) \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ r(p-1) & r(p-2) & r(p-3) & \cdots & r(0) \end{bmatrix},$$

$$\mathbf{a}^{\mathrm{T}} = [a_1 \; a_2 \; a_3 \; \cdots \; a_p],$$

where $\mathbf{a}^{\mathrm{T}}$ is the transpose of $\mathbf{a}$, and

$$\mathbf{r}^{\mathrm{T}} = [r(1) \; r(2) \; r(3) \; \cdots \; r(p)],$$

where $\mathbf{r}^{\mathrm{T}}$ is the transpose of $\mathbf{r}$.

Thus, the LPC coefficients can be obtained by solving

$$\mathbf{a} = \mathbf{R}^{-1}\mathbf{r}, \tag{10}$$

where $\mathbf{R}^{-1}$ is the inverse of $\mathbf{R}$.

Thus, we have introduced an approach for extracting a spectral feature of microarray gene expression data, which will be used for data classification. We will discuss another kind of spectral features for microarray gene expression data in the following subsection.

3.2. LPC cepstral coefficients

If we can determine the LPC for a biological sequence s_l, then we can also extract another feature as the cepstral coefficients, c_m, which are directly derived from the LPC coefficients. The LPC cepstral coefficients can be determined by the following recursion [41].

$$c_0 = \ln(G^2), \tag{11}$$

$$c_m = a_m + \sum_{k=1}^{m-1} \left(\frac{k}{m} \right) c_k a_{m-k}, \quad 1 \leqslant m \leqslant p, \tag{12}$$

$$c_m = \sum_{k=1}^{m-1} \left(\frac{k}{m} \right) c_k a_{m-k}, \quad m > p, \tag{13}$$

where G is the LPC gain, whose squared term is given as [42]

$$G^2 = r(0) - \sum_{k=1}^{p} a_k r(k). \tag{14}$$

4. Spectral distortion measures

Methods for measuring similarity or dissimilarity between two vectors or sequences is one of the most important algorithms in the field of pattern comparison and recognition. The calculation of vector similarity is based on various developments of distance and distortion measures. Before proceeding to the mathematical description of a distortion measure, we wish to point out the difference between distance and distortion functions [41], where the latter is more restricted in a mathematical sense.

Let $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{z}$ be the vectors defined on a vector space V. A metric or distance d on V is defined as a real-valued function on the Cartesian product $V \times V$ if it has the following properties

1. Positive definiteness: $0 \leqslant d(\mathbf{x}, \mathbf{y}) < \infty$, $\mathbf{x}, \mathbf{y} \in V$ and $d(\mathbf{x}, \mathbf{y}) = 0$ iff $\mathbf{x} = \mathbf{y}$;
2. Symmetry: $d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$ for $\mathbf{x}, \mathbf{y} \in V$;
3. Triangle inequality: $d(\mathbf{x}, \mathbf{z}) \leqslant d(\mathbf{x}, \mathbf{y}) + d(\mathbf{y}, \mathbf{z})$ for $\mathbf{x}, \mathbf{y}, \mathbf{z} \in V$.

If a measure of dissimilarity satisfies only the property of positive definiteness, it is referred to as a distortion measure which is considered very common for the vectorized representations of signal spectra [41]. In this sense, what we will describe next is the mathematical measure of distortion which relaxes the properties of symmetry and triangle inequality. We therefore will use the term D to denote a distortion measure. In general, to calculate a distortion measure between two vectors $\mathbf{x}$ and $\mathbf{y}$, $D(\mathbf{x}, \mathbf{y})$, is to calculate a cost of reproducing any input vector $\mathbf{x}$ as a reproduction of vector $\mathbf{y}$. Given such a distortion measure, the mismatch between two signals can be quantified by an average distortion between the input and the final reproduction. Intuitively, a match of the two patterns is good if the average distortion is small. The long-termed sample average can be expressed as [43]

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} D(\mathbf{x}_i, \mathbf{y}_i). \tag{15}$$

If the vector process is stationary and ergodic, then the limit exists and equals to the expectation of $D(\mathbf{x}_i, \mathbf{y}_i)$. Being analogous to the issue of selecting a particular distance measure for a particular problem, there is no fixed rule for selecting a distortion measure for quantifying the performance of a particular system. In general, an ideal distortion measure should be [43]:

1. tractable to allow analysis;
2. computationally efficient to allow real-time evaluation, and
3. meaningful to allow correlation with good and poor subjective quality.

To introduce the basic concept of the spectral distortion measures, we will discuss the formulation of a ratio of the prediction errors whose value can be used to expressed the magnitude of the difference between two feature vectors.

Consider passing a sequence $s(n)$ through the inverse LPC system with its LPC coefficient vector $\mathbf{a}$. This will yield the prediction error, $e(n)$, which can be alternatively defined by

$$e(n) = -\sum_{i=0}^{p} a_i s(n-i), \tag{16}$$

where $a_0 = -1$.

The sum of squared errors can be now expressed as

$$\begin{aligned} E &= \sum_{n=0}^{N-1+p} e^2(n) + \sum_{n=0}^{N-1+p} \left[-\sum_{i=0}^{p} a_i s(n-i) \right] \\ &\quad \times \left[-\sum_{j=0}^{p} a_j s(n-j) \right] \\ &= \sum_{i=0}^{p} a_i \sum_{j=0}^{p} a_j \sum_{n=0}^{N-1+p} s(n-i)s(n-j). \end{aligned} \tag{17}$$

We also have

$$\begin{aligned} \sum_{n=0}^{N-1+p} s(n-i)s(n-j) &= \sum_{n=0}^{N-1+p} s(n)s(n-j+i) \\ &= r(|i-j|). \end{aligned} \tag{18}$$

Therefore,

$$E = \sum_{i=0}^{p} a_i \sum_{j=0}^{p} a_j r(|i-j|) = \mathbf{a}^{\mathrm{T}} \mathbf{R}_s \mathbf{a}. \tag{19}$$

Similarly, consider passing another sequence $s'(n)$ through the inverse LPC system with the same LPC coefficients $\mathbf{a}$. The prediction error, $e'(n)$, is expressed as

$$e'(n) = -\sum_{i=0}^{p} a_i s'(n-i), \tag{20}$$

where $a_0 = -1$.

Using the same derivation for $s(n)$, the sum of squared errors for $s'(n)$ is

$$E' = \sum_{i=0}^{p} a_i \sum_{j=0}^{p} a_j r'(|i-j|) = \mathbf{a}^{\mathrm{T}} \mathbf{R}_{s'} \mathbf{a}, \tag{21}$$

where

$$\mathbf{R}_{s'} = \begin{bmatrix} r'(0) & r'(1) & r'(2) & \cdots & r'(p-1) \\ r'(1) & r'(0) & r'(1) & \cdots & r'(p-2) \\ r'(2) & r'(1) & r'(0) & \cdots & r'(p-3) \\ . & . & . & \cdots & . \\ r'(p-1) & r'(p-2) & r'(p-3) & \cdots & r'(0) \end{bmatrix}.$$

It can be seen that E' must be greater than or equal to E because E is the minimum prediction error for the LPC system with the LPC coefficients $\mathbf{a}$. Thus, the ratio of the two prediction errors, denoted as D, can be now defined by

$$D = \frac{E'}{E} = \frac{\mathbf{a}^{\mathrm{T}}\mathbf{R}_{s'}\mathbf{a}}{\mathbf{a}^{\mathrm{T}}\mathbf{R}_{s}\mathbf{a}} \geqslant 1. \tag{22}$$

By now it can be seen that the derivation of the above distortion is based on the concept of the *error matching measure*.

4.1. LPC likelihood distortion

Consider the two spectra, magnitude-squared Fourier transforms, $S(\omega)$ and $S'(\omega)$ of the two signals s and s', where ω is the normalized frequency ranging from $-\pi$ to π. The log spectral difference between the two spectra is defined by [41]

$$V(\omega) = \log S(\omega) - \log S'(\omega) \tag{23}$$

which is the basis for the distortion measure proposed by Itakura and Saito in their formulation of linear prediction as an approximate maximum likelihood estimation.

The Itakura–Saito distortion measure, D_{IS}, is defined as [44]

$$\begin{aligned} D_{IS} &= \int_{-\pi}^{\pi} \left[e^{V(\omega)} - V(\omega) - 1 \right] \frac{\mathrm{d}\omega}{2\pi} \\ &= \int_{-\pi}^{\pi} \frac{S(\omega)}{S'(\omega)} \frac{\mathrm{d}\omega}{2\pi} - \log \frac{\sigma_{\infty}^2}{\sigma_{\infty}'^2} - 1, \end{aligned} \tag{24}$$

where σ_{∞}^2 and $\sigma_{\infty}'^2$ are the one-step prediction errors of $S(\omega)$ and $S'(\omega)$, respectively, and defined as

$$\sigma_{\infty}^2 \approx \exp\left\{ \int_{-\pi}^{\pi} \log S(\omega) \frac{\mathrm{d}\omega}{2\pi} \right\}. \tag{25}$$

It was pointed out that the Itakura–Saito distortion measure is connected with many statistical and information theories [41] including the likelihood ratio test, discrimination information, and Kullback–Leibler divergence. Based on the notion of the Itakura–Saito distortion measure, the LPC likelihood ratio distortion between two signals s and s' is derived and expressed as [41]

$$D_{LR} = \frac{\mathbf{a}'^{\mathrm{T}}\mathbf{R}_{s}\mathbf{a}'}{\mathbf{a}^{\mathrm{T}}\mathbf{R}_{s}\mathbf{a}} - 1, \tag{26}$$

where $\mathbf{R}_s$ is the autocorrelation matrix of sequence s associated with its LPC coefficient vector $\mathbf{a}$, and $\mathbf{a}'$ is the LPC coefficient vector of signal s'.

4.2. LPC cepstral distortion

Let $S(\omega)$ be the power spectrum of a signal. The complex cepstrum of the signal is defined as the Fourier transform of the log of the signal spectrum

$$\log S(\omega) = \sum_{n=-\infty}^{\infty} c_n \mathrm{e}^{-\mathrm{j}n\omega}, \tag{27}$$

where $c_n = -c_n$ are real and referred to as the cepstral coefficients.

Consider $S(\omega)$ and $S'(\omega)$ to be the power spectra of the two signals and apply the Parseval's theorem [48], the L_2-norm cepstral distance between $S(\omega)$ and $S'(\omega)$ can be related to the root-mean-square log spectral distance as [41]

$$\begin{aligned} D_c^2 &= \int_{-\pi}^{\pi} |\log S(\omega) - \log S'(\omega)|^2 \frac{\mathrm{d}\omega}{2\pi} \\ &= \sum_{n=-\infty}^{\infty} (c_n - c_n')^2, \end{aligned} \tag{28}$$

where c_n and c_n' are the cepstral coefficients of $S(\omega)$ and $S'(\omega)$, respectively.

Since the cepstrum is a decaying sequence, the infinite number of terms in (28) can be truncated to some finite number $L \geqslant p$, that is

$$D_c^2(L) = \sum_{m=1}^{L} (c_m - c_m'). \tag{29}$$

5. Experiments

5.1. Experiment #1

In this study, the DNA sequences used for testing the proposed approach are the thrA, thrB and thrC genes of the threonine operons from *E. coli* K-12 and from *S. flexneri*; and one random sequence. In addition, we compared all six sequences with a randomly generated sequence (rand-thrA). These sequences are described in Section A.1 of Appendix.

We designed two experiments to test and compare the proposed method with other existing approaches. The first test was carried out to find out the phylogenetics between the thrA, thrB and thrC genes of the threonine operons from *E. coli* K-12 and from *S. flexneri*; and one random sequence.

To compare our proposed technique with other methods, we calculated the sequence similarity or sequence distance using alignment-based methods. All seven sequences have been aligned using CLUSTALW [13]. The multiple sequence alignment has then been used to calculate an identity matrix and the distance matrix using DNADist from the PHYLIP package [45] and the modification of the Kimura distance model [46]. The DNADist program uses nucleotide sequences to compute a distance matrix, under the modified Kimura model of nucleotide substitution. Being similar

to the Jukes and Cantor model [47], which constructs the transition probability matrix based on the assumption that a base change is independent of its identity, the Kimura 2-parameter model allows for a difference between transition and transversion rates in the construction of the DNA distance matrix.

The results obtained using all the presented spectral distortion measures agree with the SimMM [27] and the chaos game representation [26] even though we used seven sequences as test sets; where *ec*-thrA is closer to *ec*-thrC than to *ec*-thrB, and *ec*-thrB is closer to *ec*-thrA than to *ec*-thrC. This relationship was found within both species, *E. coli* K-12 and *S. flexneri*. We need to point out that this agreement between these models does not confirm any hypothesis about the relationships of these threonine operons since we have found no current phylogenetic study of these threonine operons in the literature. The alignment-based methods, on the other hand, show a slightly different relationship between the three different sequences. The calculations from both the identity and distance matrices place the thrA sequences closer to thrB than to thrC, and thrB closer to thrC than to thrA. However, the identity-matrix based model places randthrA closer to the two thrA sequences, whose relationship is not supposed to be so.

Table 2
Protein sequence specification

Protein	Size	GenBankID
(HAHU) Human α hemoglobin	142	HAHU
(HAHO) Horse α hemoglobin	142	HAHO
(HBHU) Human β hemoglobin	147	HBHU
(CCPG) Pig cytochrome *c*	104	CCPG
(LEGH) Lupine leghemoglobin	154	P02239
(LZRT) Rat lysozome	148	LZRT
(MYWHP) Sperm whale myoglobin	153	MYWHP
(FGFB) Basic human Fibroblast growth factors	288	NP_001997.4

Table 3
Structural specification of 8 proteins

Protein	Class	Super family	Family
HAHU	All α proteins	Globin-like	Globin
HAHO	All α proteins	Globin-like	Globin
HBHU	All α proteins	Globin-like	Globin
CCPG	All α proteins	Cytochrome *c*	Cytochrome *c*
LEGH	All α proteins	Globin-like	Globin
LZRT	α and β proteins	Lysozyme-like	C-type lysozyme
MYWHP	All α proteins	Globin-like	Globin
FGFB	All β proteins	Cytokine	Fibroblast growth factors (FGF)

5.2. Experiment #2

The second test involves a database of 40 complex DNA sequences, which was used for searching similar sequences to a query sequence. Description of the sequences is given in Section A.2 of Appendix.

The proposed spectral distortion measures were further tested to search for DNA sequences being similar to a query sequence from a database of 39 library sequences, of which 20 sequences are known to be similar in biological function to the query sequence, and the remaining 19 sequences are known as being not similar in biological function to the query sequence. These 39 sequences were selected from mammals, viruses, plants, etc., of which lengths vary between 322 and 14 121 bases. The query sequence is HSLIPAS (Human mRNA for lipoprotein lipase).

Sensitivity and selectivity were computed to evaluate and compare the performance of the proposed models with other distance measures [23]. Sensitivity is expressed by the number of HSLIPAS related sequences found among the first closest 20 library sequences; whereas selectivity is expressed in terms of the number of HSLIPAS-related sequences of which distances are closer to HSLIPAS than others and are not truncated by the first HSLIPAS-unrelated sequence. Among several distance measures introduced by Wu et al. [23], they concluded that the standardized Euclidean distance under the Markov chain models of base composition was generally recommended, of which sensitivity and selectivity are 18 and 17 sequences respectively, of order one for base composition, and 18 and 16 sequences, respectively, of order two for base composition; when all the distances of nine different word sizes were combined. From our previous study, both sensitivity and selectivity obtained from SimMM [27] are 18 sequences. The sensitivity and selectivity obtained from the LPC likelihood distortion are 19 and 18 sequences, respectively; whereas the LPC cepstral distortion achieved 20 sequences for both sensitivity and selectivity. The results obtained from the distortion measures show their superiority over the other methods for database searching of similar DNA sequences.

5.3. Experiment #3

To test the performance of our proposed approach with protein sequences, we selected the same protein data sets which were studied by Trad et al. [31]. The protein sequences with their GenBank identities are shown in Table 2, and described in Appendix B (Table 3).

Tables 4–6 show the similarities of the eight protein sequences presented in Table 2 by a truncated LPC cepstrum distortion measure with $p = 3$ and $L = 6$, 9, and 18, respectively. It can be observed that the distortion measures from different values of L are not much different from one another. We choose the results obtained from $L = 6$ for comparing with the results obtained from a wavelet-based protein sequence comparison [31], which applied the EIIP values.

Table 7 shows the result of wavelet based method with the same eight protein sequences. In the wavelet-based protein sequence comparison, the authors used the term S to denote

Table 4
Similarity matrix by cepstral-distortion, $p = 3$ and $L = 6$

	HAHU	HAHO	HBHU	CCPG	LEGH	LZRT	MYWHP	FGFH
HAHU	0	0.0029	0.0006	0.0582	0.0148	0.0050	0.0023	0.0512
HAHO	0.0029	0	0.0043	0.0750	0.0169	0.0099	0.0042	0.0546
HBHU	0.0006	0.0043	0	0.0488	0.0133	0.0081	0.0033	0.0479
CCPG	0.0582	0.0750	0.0488	0	0.0365	0.0859	0.0562	0.0384
LEGH	0.0148	0.0169	0.0133	0.0365	0	0.0331	0.0072	0.0112
LZRT	0.0050	0.0099	0.0081	0.0859	0.0331	0	0.0099	0.0793
MYWHP	0.0023	0.0042	0.0033	0.0562	0.0072	0.0099	0	0.0350
FGFH	0.0512	0.0546	0.0479	0.0384	0.0112	0.0793	0.0350	0

Table 5
Similarity matrix by cepstral-distortion, $p = 3$ and $L = 9$

	HAHU	HAHO	HBHU	CCPG	LEGH	LZRT	MYWHP	FGFH
HAHU	0	0.0031	0.0006	0.0582	0.0149	0.0053	0.0023	0.0515
HAHO	0.0031	0	0.0044	0.0752	0.0172	0.0109	0.0046	0.0555
HBHU	0.006	0.0044	0	0.0488	0.0134	0.0086	0.0034	0.0483
CCPG	0.0582	0.0752	0.0488	0	0.0366	0.0862	0.0563	0.0387
LEGH	0.0149	0.0172	0.0134	0.0366	0	0.0333	0.0072	0.0113
LZRT	0.0053	0.0109	0.0086	0.0862	0.0333	0	0.0100	0.0793
MYWHP	0.0023	0.0046	0.0034	0.0563	0.0072	0.0100	0	0.0351
FGFH	0.0515	0.0555	0.0483	0.0387	0.0113	0.0793	0.0351	0

Table 6
Similarity matrix by cepstral-distortion, $p = 3$ and $L = 18$

	HAHU	HAHO	HBHU	CCPG	LEGH	LZRT	MYWHP	FGFH
HAHU	0	0.0032	0.0006	0.0582	0.0149	0.0056	0.0024	0.0517
HAHO	0.0032	0	0.0045	0.0752	0.0174	0.0117	0.0049	0.0562
HBHU	0.0006	0.0045	0	0.0488	0.0134	0.0090	0.0035	0.0486
CCPG	0.0582	0.0752	0.0488	0	0.0366	0.0866	0.0564	0.0390
LEGH	0.0149	0.0174	0.0134	0.0366	0	0.0335	0.0072	0.0115
LZRT	0.0056	0.0117	0.0090	0.0866	0.0335	0	0.0101	0.0793
MYWHP	0.0024	0.0049	0.0035	0.0564	0.0072	0.0101	0	0.0352
FGFH	0.0517	0.0562	0.0486	0.0390	0.0115	0.0793	0.0352	0

Table 7
Similarity by wavelet-based method [31]

	HAHU	HAHO	HBHU	CCPG	LEGH	LZRT	MYWHP	FGFH
HAHU	5S	5S	1S1W3N	1W4N	2W3N	5N	2W3N	5N
HAHO	5S	5S	1S1W3N	1W4N	2W3N	5N	W3N	5N
HBHU	1S1W3N	1S1W3N	5S	1W4N	2W3N	5N	2W3N	5N
CCPG	1W4N	1W4N	1W4N	5S	1W4N	1W4N	5N	5N
LEGH	2W3N	2W3N	2W3N	1W4N	5S	1W4N	1W3N	5N
LZRT	5N	5N	5N	1W4N	1W4N	5S	5N	1W4N
MYWHP	2W3N	2W3N	2W3N	5N	1W3N	5N	5S	1W4N
FGFH	5N	5N	5N	5N	5N	1W4N	1W4N	5S

Table 8
Structural classification of 40 protein sequences

Index	ID	Class	Fold	Superfamily	Family
Query	HAHU	All alpha proteins	Globin-like		Globins
1	HAHO	All alpha proteins	Globin-like		Globins
2	HBHU	All alpha proteins	Globin-like		Globins
3	LEGH	All alpha proteins	Globin-like		Globins
4	MYWHP	All alpha proteins	Globin-like		Globins
5	MYHO	All alpha proteins	Globin-like		Globins
6	P01966	All alpha proteins	Globin-like		Globins
7	HAMS	All alpha proteins	Globin-like		Globins
8	1ECO	All alpha proteins	Globin-like		Globins
9	P06148	All alpha proteins	Globin-like		Globins
10	P39662	All alpha proteins	Globin-like		Globins
11	P24232	All alpha proteins	Globin-like		Globins
12	P02207	All alpha proteins	Globin-like		Globins
13	P04252	All alpha proteins	Globin-like		Globins
14	1HLB	All alpha proteins	Globin-like		Globins
15	CCPG	All alpha proteins	Cytochrome c		Cytochrome c
16	LZRT	Alpha and beta proteins	Lysozyme-like		C-type lysozyme
17	FGFH	All beta proteins	beta-Trefoil	Cytokine	Fibroblast growth factors (FGF)
18	LZCH	Alpha and beta proteins	Lysozyme-like		C-type lysozyme
19	P81708	Alpha and beta proteins	Lysozyme-like		C-type lysozyme
20	1HFY	Alpha and beta proteins	Lysozyme-like		C-type lysozyme
21	1HFX	Alpha and beta proteins	Lysozyme-like		C-type lysozyme
22	AAK54734	All beta proteins	PEBP-like		Phosphatidylethanolamine binding protein
23	P13696	All beta proteins	PEBP-like		Phosphatidylethanolaminebinding protein
24	NP_083871	All beta proteins	PEBP-like		Phosphatidylethanolamine binding protein
25	Q41261	All beta proteins	PEBP-like		Phosphatidylethanolamine binding protein
26	P29965	All beta proteins	TNF-like		TNF-like
27	NP_690616	All beta proteins	TNF-like		TNF-like
28	NP_776480	All beta proteins	beta-Trefoil	Cytokine	Fibroblast growth factors (FGF)
29	NP_071518	All beta proteins	beta-Trefoil	Cytokine	Fibroblast growth factors (FGF)
30	P05619	Multi-domain proteins			Serpins
31	P01012	Multi-domain proteins			Serpins
32	P01011	Multi-domain proteins			Serpins
33	P41361	Multi-domain proteins			Serpins
34	O35684	Multi-domain proteins			Serpins
35	P14754	Multi-domain proteins			Serpins
36	P13299	Multi-domain proteins			DNA-binding domain of intron endonuclease I-TevI
37	Q9P9H1	Multi-domain proteins			DNA primase
38	1AD2	Multi-domainproteins			Ribosomal protein L1
39	P54050	Multi-domain proteins			Ribosomal protein L1
40	P06179	Multi-domain proteins			F41 fragment of flagellin

a strong correlation (> 0.7), W to denote a weak correlation (0.5 to 0.7), and N to denote no correlation (< 0.5).

For the wavelet-based approach, the cross-correlation of HAHU and HAHO is 5S (highly correlated); between HAHU and HBHU the cross-correlation is 1S1W3N; and the cross-correlation of LEGH and MYWHP against HAHU are 2W3N. Those sequences belong to the same protein family. Even though LEGH has only 14% identical residues with HAHU, they are distantly related. Two weak correlations are detected in two different resolutions by the wavelet-based method.

Due to the fact that lysozyme and hemoglobin do not share any biological function, we therefore chose the lowest value, which is 0.005, of the distortion measures between LZRT and HAHU, HAHO, HBHU, LEGH, MWHP to establish the threshold for identifying similar sequences. Our approach detects HAHU, HAHO, HBHU and MYWHP as related proteins by the cutoff value. The distances from HAHU to HAHO, HBHU and MYWHP are 0.0029, 0.0006 and 0.0023, respectively. These values are well below the cutoff value.

5.4. Experiment # 4

In this experiment, we used 40 protein sequences (HAHU: query sequence, 7 pre-experimented sequences, and 33 additional sequences) to test our proposed method in this second experiment. Table 8 presents the structural classification of query sequence (HAHU) and 40 protein

Table 9
Protein sequence identities from alignment-based analysis

Index	ID	Size	Identity (%)
Query	HAHU	142	
1	HAHO	142	88.0
2	HBHU	147	43.4
3	LEGH	154	20.3
4	MYWHP	153	26.0
5	MYHO	153	26.2
6	P01966	142	88.0
7	HAMS	142	85.9
8	1ECO	136	22.2
9	P06148	142	72.7
10	P39662	403	26.7
11	P24232	396	25.3
12	P02207	149	31.8
13	P04252	146	21.2
14	1HLB	158	22.4
15	CCPG	104	26.4
16	LZRT	148	29.4
17	FGFH	288	22.2
18	LZCH	147	28.6
19	P81708	129	32.5
20	1HFY	123	83.3
21	1HFX	123	35.7
22	AAK54734	104	40.0
23	P13696	187	21.6
24	NP_083871	192	21.8
25	Q41261	181	23.1
26	P29965	261	31.4
27	NP_690616	103	60.0
28	NP_776480	155	29.2
29	NP_071518	194	25.4
30	P05619	379	23.9
31	P01012	386	17.4
32	P01011	423	17.5
33	P41361	433	26.0
34	O35684	410	23.9
35	P14754	392	35.3
36	P13299	245	19.5
37	Q9P9H1	347	40.0
38	1AD2	228	26.0
39	P54050	219	38.9
40	P06179	495	22.9

Table 10
Similarity measures by LPC cepstral-distortion measure

Index	ID	Family	Similarity
Query	HAHU	Globins	0.0000
1	HAHO	Globins	0.0031
2	HBHU	Globins	0.0006
3	LEGH	Globins	0.0149
4	MYWHP	Globins	0.0023
5	MYHO	Globins	0.0014
6	P01966	Globins	0.0035
7	HAMS	Globins	0.0042
8	1ECO	Globins	0.0007
9	P06148	Globins	0.0037
10	P39662	Globins	0.0043
11	P24232	Globins	0.0030
12	P02207	Globins	0.0078
13	P04252	Globins	0.0524
14	1HLB	Globins	0.0150
15	CCPG	Cytochrome c	0.0582
16	LZRT	C-type lysozyme	0.0053
17	FGFH	Fibroblast growth factors (FGF)	0.0515
18	LZCH	C-type lysozyme	0.0201
19	P81708	C-type lysozyme	0.0052
20	1HFY	C-type lysozyme	0.0610
21	1HFX	C-type lysozyme	0.0318
22	AAK54734	Phosphatidylethanolamine binding protein	0.0081
23	P13696	Phosphatidylethanolamine binding protein	0.0028
24	NP_083871	Phosphatidylethanolamine binding protein	0.0065
25	Q41261	Phosphatidylethanolamine binding protein	0.0503
26	P29965	TNF-like	0.0080
27	NP_690616	TNF-like	0.0082
28	NP_776480	Fibroblast growth factors (FGF)	0.0638
29	NP_071518	Fibroblast growth factors (FGF)	0.0403
30	P05619	Serpins	0.0076
31	P01012	Serpins	0.0047
32	P01011	Serpins	0.0150
33	P41361	Serpins	0.0048
34	O35684	Serpins	0.0096
35	P14754	Serpins	0.0055
36	P13299	DNA-binding domain of intron endonuclease I-TevI	0.0031
37	Q9P9H1	DNA primase	0.0044
38	1AD2	Ribosomal protein L1	0.0139
39	P54050	Ribosomal protein L1	0.0153
40	P06179	F41 fragment of flagellin	0.0472

sequences from 4 different protein classes, 12 different folds, 12 different superfamilies and 12 different protein families. This structural classification is based on SCOP (Structural Classification of Proteins), being available at http://scop.mrclmb.com.ac.uk/scop, which classifies proteins by class, fold, superfamily, and family.

Similar to the first experiment, the query sequence was also HAHU (α hemoglobin) which belongs to all α proteins class, globin-like fold, globin-like superclass, and globins family. Among the 40 sequences, sequences 1 to 14 belong to the same class, fold, superfamily and family as the query sequence (HAHU). They all belong to the globins family. The remainders, from sequences 15 to 40, belong to 11 different protein families. Therefore, the criterion of this experiment was that the distances between query sequence (HAHU) and 14 globins family proteins (from index numbers 1 to 14 proteins in Table 8) should be closer than any other proteins of different groups.

Table 9 shows the sequence identities between the query sequence (HAHU) and other 40 protein sequences obtained by the optimal local alignment-based method (Smith–Waterman algorithm). The results of our proposed method are shown in Table 10, where truncated LPC cepstrum distortion measure was used with $p = 3$, and $L = 9$.

Sensitivity is defined as the top 14 sequences being most similar to the query sequence. The local alignment-based

Table 11
Sensitivity and selectivity by local alignment and cepstral-distortion methods

	Local alignment	Cepstral distortion
Sensitivity	5/14	10/14
Selectivity	3/14	4/14

methods detected 5 globin proteins among 14 globins proteins. On the other hand, our proposed method, LPC cepstrum based similarity measure, detected 10 globin proteins out of 14. Table 11 presents the sensitivity obtained by the alignment-based method and our proposed method. The LPC cepstral based method shows better performance than the alignment method.

Selectivity is expressed in terms of the number of query-related sequences of which distances are closer to the query sequence (HAHU) than others and are not truncated by the first HAHU-unrelated sequence. The local alignment-based method obtained 3 out of 14 related sequences. For the local alignment method, the identity of protein sequence number 20 (1HFY), which belong to α and β proteins class, lysozyme-like fold and superfamily, and *c*-type lysozyme family, is 83.3%. However, this protein is from a totally different class, fold, superfamily and family. The LPC cepstral based method obtained 4 out of 14 related sequences. Our proposed method shows a better performance than the local alignment method. Table 11 also shows the comparison in terms of the selectivity between the LPC cepstral distortion measure and the alignment-based method.

6. Conclusions

We have introduced the LPC cepstral-distortion measure for the analysis of protein sequences. The experimental results and the comparisons have shown the effectiveness of the proposed approach. The model is both physically reasonable and mathematically tractable. In speech recognition [53] as well as this study, the performance of the LPC cepstral distortion measure appears to perform better than that of the LPC likelihood distortion measure. Moreover the framework for computing the LPC-based cepstral distortion is very efficient for real-time computer implementation and comparison of protein signals.

Regarding the computational considerations of the proposed model, the solution of the LPC cepstral distortion measure mainly involves in the computations of the LPC autocorrelation and the truncated cepstral distortion. For the autocorrelation method, to obtain the LPC coefficients, the p linear equations in (9) need to be solved, which require the inversion of the $\mathbf{R}$ matrix and the multiplication of the resultant $p \times p$ matrix with the $\mathbf{r}$ vector. However, the redundancy in the $\mathbf{R}$ matrix allows efficient computation of (10) without resorting to the explicit inverse of the $p \times p$ matrix $\mathbf{R}$. Firstly $\mathbf{R}$ is symmetric, and secondly $\mathbf{R}$ is also a Toeplitz matrix in which all diagonal elements are equal. The autocorrelation method requires only $2p$ storage locations and $O(p^2)$ operations. Therefore, a solution for the LPC coefficients can be solved recursively and most effectively using the Levinson–Durbin algorithm [40] without operating on matrix inversion.

In general, LPC and its distortion measures can be a rigorous approach for solving the problem of pattern comparisons of biological sequences and worth further investigation. In one aspect, results of the proposed method greatly depend on the EIIP values. Thus, distortion measures for DNA and protein sequences analysis can certainly be improved if more effective ways for numerical representation of the sequences can be explored.

Acknowledgements

The author acknowledges the assistance of Byung–Sub Shim who provided the information on the protein sequences and carried out the computer implementation of the proposed methods under the supervision of the author.

Appendix A. DNA data

A.1. Threonine operons

The data are taken from the threonine operons of *E. coli* K-12 (gi:1786181) and *S. flexneri* (gi:30039813). The three sequences taken from each threonine operon are thrA (aspartokinase I-homoserine dehydrogenase I), thrB (homoserine kinase) and thrC (threonine synthase), using the open reading frames (ORFs) 3372799 (*ec*-thrA), 28013733 (*ec*-thrB) and 37345020 (*ec*-thrC) in the case of *E.coli* K-12, and 336-2798 (*sf*-thrA), 28003732 (*sf*-thrB) and 37335019 (*sf*-thrC) in the case of *S. flexneri*. All the sequences were obtained from GenBank (www.ncbi.nlm.nih.gov/Entrez). We generated a random sequence (rand-thrA), using the same length and base composition as *ec*-thrA.

A.2. Database

The database contains 39 library sequences. The query sequence, HSLIPAS (Human mRNA for lipoprotein lipase), has 1612 bases. All of these sequences can be obtained from the GenBank sequence database (http://www.ncbi.nlm.nih.gov/Entrez/).

The 20 sequences, which are known as being similar in biological function to HSLIPAS are as follows: OOLPLIP (*Oestrus ovis* mRNA for lipoprotein lipase, 1656 bp), SSLPLRNA (pig back fat Sus scrofa cDNAsimilar to S.scrofa LPL mRNA for lipoprotein lipase, 2963 bp), RATLLIPA (*Rattus norvegicus* lipoprotein lipase mRNA, complete cds, 3617 bp), MUSLIPLIP (*Mus musculus*

lipoprotein lipase gene, partial cds, 3806 bp), GPILPPL (guinea pig lipoprotein lipase mRNA, complete cds, 1744 bp), GGLPL (chicken mRNA for adipose lipoprotein lipase, 2328 bp), HSHTGL (human mRNA for hepatic triglyceride lipase, 1603 bp), HUMLIPH (human hepatic lipase mRNA, complete cds, 1550 bp), HUMLIPH06 (human hepatic lipase gene, exon 6, 322 bp), RATHLP (rat hepatic lipase mRNA, 1639 bp), RABTRIL [*Oryctolagus cuniculus* (clone TGL-5K) triglyceride lipase mRNA, complete cds, 1444 bp], ECPL (*Equus caballus* mRNA for pancreatic lipase, 1443 bp), DOGPLIP (canine lipase mRNA, complete cds, 1493 bp), DMYOLK [Drosophila gene for yolk protein I (vitellogenin), 1723 bp], BOVLDLR [bovine low-density lipoprotein (LDL) receptor mRNA, 879 bp], HSBMHSP (Homosapiens mRNA for basement membrane heparan sulfate proteoglycan, 13 790 bp), HUMAPOAICI (human apolipoprotein A-I and C-III genes, complete cds, 8966 bp), RABVLDLR (*O.cuniculus* mRNA for very LDL receptor, complete cds, 3209 bp), HSLDL100 (human mRNA for apolipoprotein B-100, 14 121 bp) and HUMAPOBF (human apolipoprotein B-100 mRNA, complete cds, 10 089 bp).

The other 19 sequences known as being not similar in biological function to HSLIPAS are as follows: A1MVRNA2 [*alfalfa* mosaic virus (A1M4) RNA 2, 2593 bp], AAHAV33A [*Acanthocheilonema viteae* pepsin-inhibitorlike-protein (Av33) mRNA sequence, 1048 bp], AA2CG (adeno-associated virus 2, complete genome, 4675 bp), ACVPBD64 (artificial cloning vector plasmid BD64, 4780 bp), AL3HP (bacteriophage alpha-3 H protein gene, complete cds, 1786 bp), AAABDA[Aedes aegypti abd-A gene for abdominal-A protein homolog (partial), 1759 bp], BACBDGALA [*Bacillus circulans* beta-d-galactosidase (bgaA) gene, complete cds, 2555 bp], BBCA (Bos taurus mRNA for cyclin A, 1512 bp), BCP1 (bacteriophage Chp1 genome DNA, complete sequence, 4877 bp) and CHIBATPB(sweet potato chloroplast F1-ATPase beta and epsilon-subunit genes, 2007 bp), A7NIFH (*Anabaena* 7120 nifH gene, complete CDS, 1271 bp), AA16S (*Amycolatopsis azurea* 16S rRNA, 1300 bp), ABGACT2 (*Absidia glauca* actin mRNA, complete cds, 1309 bp), ACTIBETLC (Actinomadura R39 DNA for beta-lactamase gene, 1902 bp), AMTUGSNRNA (*Ambystoma mexicanum* AmU1 snRNA gene, complete sequence, 1027 bp), ARAST18B (cloning vector pAST 18b for *Caenorhabditis elegans*, 3052 bp), GCALIP2 (*Geotrichum candidum* mRNA for lipase II precursor, partial cds, 1767 bp), AGGGLINE (*Ateles geoffroyi* gamma-globin gene and L1 LINE element, 7360 bp) and HUMCAN (*H.sapiens* CaN19 mRNA sequence, 427 bp).

Appendix B. Protein data

- *Hemoglobin*: Hemoglobin is the iron-containing oxygen-transport metalloprotein in the red cells of the blood in mammals and other animals. It transports oxygen to tissues. It is one of the most well-known globin proteins. The structural point of view, hemoglobin is a tetrameric molecule whose quaternary structure is comprised of two α and two β peptide chains. The subunits are structurally similar and also the size of subunits quite similar, and each subunit of hemoglobin contains one heme, therefore, each hemoglobin can bind four oxygen. The name hemoglobin is the concatenation of heme and globin, because of oxygen-binding proteins are called as a globin protein and each subunit of hemoglobin have one heme. Although the sequences of α hemoglobin and β hemoglobin are not completely identical, they have exactly the same biological function [49].
- *Myoglobin*: Myoglobin is a single-chain molecule, which is containing a heme group. It is also an oxygen-carrying protein like hemoglobin. The main role of this molecule is oxygen-carrying to muscle tissues. The myoglobin molecule is built up of eight helices, which compose a box-like structure with a hydrophobic pocket. The heme group responsible for oxygen binding is fixed in this pocket only by weak bonding [31]. It is structurally related, however, this protein does not exhibit cooperative binding of oxygen. Instead, the binding of oxygen by myoglobin is unaffected by the oxygen tension in the surrounding tissue. Obviously, myoglobin also belong to globin family, and it has the highest oxygen affinity compared with other globin proteins, such as hemoglobin, this means that other globin proteins releases oxygen to tissues more easily than myoglobin does [50].
- *Leghemoglobin*: Leghemoglobin is an iron-containing, hemoglobin-like oxygen binding red pigment(s) produced in root nodules during the symbiotic association between Bradyrhizobium or Rhizobium and legumes. The pigment resembles but is not identical to mammalian hemoglobin The red pigment has a molecular weight approximately 1/4 that of hemoglobin and has been suggested to act as an oxido-reduction catalyst in symbiotic nitrogen fixation.
- *Cytochrome c*: Cytochrome *c* is another heme-containing protein, which is soluble protein, unlike other cytochromes. It transfers electrons from the QH2-cytochrome *c* reductase complex to the cytochrome *c* oxidase complex [31]. It transfers electrons between Complexes III and IV. Moreover, cytochrome *c* is a highly conserved protein across the spectrum of species, found in plants, animals, and many unicellular organisms. And the size of this protein relatively small which make it handy to study of evolutionary divergence. Cytochrome *c* is capable of undergoing oxidation and reduction, However, it does not bind oxygen. It is one of major difference from globin family proteins.
- *Lysozyme*: Lysozyme is an enzyme. In general, it is referred to as the body's own antibiotic. It is abundantly present in a number of secretions, such as tears. This protein is presented in cytoplasmic granules of the polymorphonuclear neutrophils and released to the mucosal secretions. Lysozyme levels in the blood are often in-

creased in sarcoidosis. Basically, this enzyme does not share any biological function with globin proteins.

- *Fibroblast growth factors*: Fibroblast growth factors (FGFs) are pleiotropic mitogenic activators thlammatory cylly resemble the IL-1 family of inflammatory cytokines. Acidic FGF (aFGF) and basic FGF(bFGF) bind negatively charged heparin-like molecules, as dose VEGF. FGFs play multiple roles as embryonic inducers, endothelial mitogens, and stimulators of protease activity. FGFs may also be trophic for glial cells and neurons, and can facilitate hemopoiesis Keratinocyte growth factor is distinguished from other FGFs by its target specificity for epithelial cells and lack of mitogenicity for fibroblasts [50]. FGFs are unrelated to globin family proteins.

In terms of structural classification, human α hemoglobin (HAHU), horse α hemoglobin (HAHO), human β hemoglobin (HBHU), Lupine leghemoglobin (LEGH) and sperm whale myoglobin (MYWHP) belong to a same super family, and others to different super family. Those sequences have heme (or hame) group in common, and There are many different hemes. Three kinds of heme are considered as biologically important kinds of heme: heme *a*, heme *b*, the most common type, and heme *c*. HAHU, HAHO, HBHU, LEGH and MYWHP are of heme *a*; CCPG is of heme *c*. Based on the protein structure, HAHU, HAHO, HBHU, LEGH, MYWHP are more similar than others, and even though CCPG belongs to cytocrome *c* super family, it belongs to the same class as globin-like proteins so it should be the next closest. The above five sequences, which are in globin-like super family, are called heme protein, are used to share oxygen binding function. Cytocrome *c* is of the heme group; however, it does not hold oxygen and it has different type of heme. This is the major biological difference between cytocrome *c* and globin-like proteins.

Human α hemoglobin (HAHU) and horse α hemoglobin (HAHO) are orthologous sequences. Orthologous proteins are found in two or more organisms that have very high similarity that they almost have similar three-dimensional structure, domain structure, and biological function [51]. Paralogous sequences have been known to be related through gene duplication events. These events may lead to the production of a family of related proteins with similar sequences but also variable biological functions within a species [51]. HAHU and HBHU are paralogous sequences. Those sequences have about 43% of identical residues. The similarity of sperm whale myoglobin and lupine leghemoglobin have only 19% identical residues. However, those sequences belong to the same family, and the same class. They share very similar biological function and have similar structure [52]. This is a typical example of distantly related protein sequences. The physical characteristics of myoglobin and hemoglobin are different in oxygen affinity and the number of chains. However, from the biological similarity standpoint, they are highly related sequences. The sequence similarity of myoglobin and hemoglobin is in the twilight zone. Those sequences have about only 24% of identical residues. Leghemoglobin and hemoglobin are distantly related protein sequences which have 14 to 20% sequence identity. Therefore, they are also difficult to detect using conventional protein sequence similarity measuring methods, even though they are clearly in the same super family and family as well. Human basic FGFs and rat lysozyme belong to different super family. Their biological function is also far from globin-like proteins. Table 3 shows the structural specification of the 8 protein sequences.

References

[1] S.B. Nagl, Function prediction from protein sequence, in: C.A. Orengo, D.T. Jones, J.M. Thornton (Eds.), Bioinformatics: Genes, Proteins and Computers, BIOS Scientific Publishers, Oxford, UK, 2003.

[2] M. Bishop, C. Rawlings, Nucleic Acid and Protein Sequence Analysis—A Practical Approach, IRL Press, Oxford, 1987.

[3] W.J. Ewens, G.R. Grant, Statistical Methods in Bioinformatics, Springer, NY, 2001.

[4] W. Miller, Comparison of genomic DNA sequences: solved and unsolved problems, Bioinformatics 17 (2001) 391–397.

[5] D.E. Krane, M.L. Raymer, Fundamental Concepts of Bioinformatics, Benjamin Cummings, San Francisco, 2003.

[6] R. Bellman, Dynamic Programming, Princeton University Press, Princeton, NJ, 1957.

[7] S.B. Needleman, C.D. Wunsch, A general method applicable to the search for similarities in the amino acid sequence of two proteins, J. Mol. Biol. 48 (1970) 443–453.

[8] T. Smith, M. Waterman, Identification of common molecular subsequences, J. Mol. Biol. 147 (1981) 195–197.

[9] R. Durbin, S.R. Eddy, A. Krogh, G.J. Mitchison, Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids, Cambridge University Press, Cambridge, UK, 1998.

[10] S. Asltschul, W. Gish, W. Miller, E. Myers, D. LIPMAN, Basic local alignment search tool, J. Mol. Biol. 215 (1990) 403–410.

[11] D.J. Lipman, W.R. Pearson, Rapid and sensitive protein similarity searches, Science 227 (1985) 1435–1441.

[12] D.G. Higgins, P.M. Sharp, CLUSTAL: a package for performing multiple sequence alignment on a microcomputer, Gene 73 (1988) 237–244.

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

[14] V. Veljkovic, I. Cosic, B. Dimitrijevic, D. Lalovic, Is it possible to analyze DNA and protein sequences by the methods of digital signal processing?, IEEE Trans. Biomed. Eng. 32 (1985) 337–341.

[15] D. Anatassiou, Frequency-domain analysis of biomolecular sequences, Bioinformatics 16 (2000) 1073–1082.

[16] P. Lio, Wavelets in bioinformatics and computational biology: state of art and perspectives, Bioinformatics 19 (2003) 2–9.

[17] L. Li, R. Jin, P.L. Kok, W. Wan, Pseudo-periodic partitions of biological sequences, Bioinformatics 20 (2004) 295–306.

[18] C.A. del Carpio-Munoz, J.C. Carbajal, Folding pattern recognition in proteins using spectral analysis methods, Genome Inform. 13 (2001) 163–172.

[19] D. Anatassiou, Genomic signal processing, IEEE Signal Process. Mag. 18 (2001) 8–20.

[20] S. Vinga, J. Almeida, Alignment-free sequence comparison—a review, Bioinformatics 19 (2003) 513–523.

[21] B.E. Blaisdell, A measure of the similarity of sets of sequences not requiring sequence alignment, Proc. Natl. Acad. Sci. USA 83 (1986) 5155–5159.
[22] T.J. Wu, J.P. Burke, D.B. Davison, A measure of DNA sequence dissimilarity based on Mahalanobis distance between frequencies of words, Biometrics 53 (1997) 1431–1439.
[23] T.J. Wu, Y.C. Hsieh, L.A. Li, Statistical measures of DNA dissimilarity under Markov chain models of base composition, Biometrics 57 (2001) 441–448.
[24] G.W. Stuart, K. Moffett, S. Baker, Integrated gene and species phylogenies from unaligned whole genome protein sequences, Bioinformatics 18 (2002) 100–108.
[25] M. Li, J.H. Badger, X. Chen, S. Kwong, P. Kearney, H. Zhang, An information-based sequence distance and its application to whole mitochondrial genome phylogeny, Bioinformatics 17 (2001) 149–154.
[26] J.S. Almeida, J.A. Carrico, A. Maretzek, P.A. Noble, M. Fletcher, Analysis of genomic sequences by chaos game representation, Bioinformatics 17 (2001) 429–437.
[27] T.D. Pham, J. Zuegg, A probabilistic measure for alignment-free sequence comparison, Bioinformatics 20 (2004) 3455–3461.
[28] I. Cosic, Macromolecular bioactivity: is it resonant interaction between macromolecules?—theory and applications, IEEE Trans. Biomed. Eng. 41 (1994) 1101–1114.
[29] V. Veljkovic, I. Slavic, General model of pseudopotentials, Phys. Rev. Lett. 29 (1972) 105–108.
[30] J. Lazovic, Selection of amino acid parameters for Fourier transform-based analysis of proteins, CABIOS 12 (1996) 553–562.
[31] C.H. de Trad, Q. Fang, I. Cosic, Protein sequence comparison based on the wavelet transform approach, Protein Eng. 15 (2002) 193–203.
[32] I. Daubechies, The wavelet transform, time-frequency localization and signal analysis, IEEE Trans. Inform. Theory 36 (1990) 961–1005.
[33] I. Daubechies, Where do wavelets come from?, Proceeding of the IEEE 84 (1996) 510–513 (Special Issue on Wavelets).
[34] Y.Y. Tang, J. Liu, L.H. Yang, Wavelet Theory and its Application to Pattern Recognition, World Scientific, Singapore, 1999.
[35] Q. Wu, K.R. Castleman, Automated chromosome classification using wavelet-based band pattern descriptors, Proceedings of 13th IEEE Symposium on Computer-Based Medical Systems, 2000, pp. 189–194.
[36] J.Z. Wang, G. Wiederhold, O. Firschein, S.X. Wei, Content-based image indexing and searching using Daubechies' wavelets, Int. J. Digital Libraries 1 (1998) 311–328.
[37] Y. Hu, P.C. Loizou, Speech enhancement based on wavelet thresholding the multitaper spectrum, IEEE Trans. Speech Audio Process. 12 (2004) 59–67.
[38] G.Y. Chen, T.D. Bui, Invariant fourier-wavelet descriptor for pattern recognition, Pattern Recognition 32 (1999) 1083–1088.
[39] E. Pirogova, G.P. Simon, I. Cosic, Investigation of the applicability of dielectic relaxation properties of amino acid solutions within the resonant recognition model, IEEE Trans. Nanobioscience 2 (2003) 63–69.
[40] J. Makhoul, Linear prediction: a tutorial review, Proc. IEEE 63 (1975) 561–580.
[41] L. Rabiner, B.H. Juang, Fundamentals of Speech Recognition, Prentice-Hall, NJ, 1993.
[42] V.K. Ingle, J.G. Proakis, Digital Signal Processing Using Matlab V.4, Boston, PWS Publishing, MA, 1997.
[43] R.M. Gray, Vector quantization, IEEE ASSP Mag. 1 (1984) 4–29.
[44] F. Itakura, S. Saito, A statistical method for estimation of speech spectral density and formant frequencies, Electron. Commun. Japan 53A (1970) 36–43.
[45] J. Felsenstein, PHYLIP (Phylogeny Inference Package), version 3.5c. Distributed by the Author, Department of Genetics, University of Washington, Seattle, WA, 1993.
[46] M. Kimura, A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences, J. Mol. Evol. 16 (1980) 111–120.
[47] T.H. Jukes, C.R. Cantor, Evolution of protein molecules, in: H.N. Munro (Ed.), Mammalian Protein Metabolism, Academic Press, NY, 1969, pp. 21–132.
[48] D. O'Shaughnessy, Speech Communication—Human and Machine, Addison–Wesley, Reading, MA, 1987.
[49] A.L. Lehninger, D.L. Nelson, M.M. Cox, Principles of Biochemistry, Worth Publishing, New York, 1993.
[50] R.J. Epstein, Human Molecular Biology: An Introduction to the Molecular Basis of Health and Disease, Cambridge University Press, Cambridge, 2003.
[51] D.W. Mount, Bioinformatics—Sequence and Genome Analysis, second ed., Cold Spring Harbor Laboratory Press, New York, 2004.
[52] R. Doolittle, Similar amino acid sequences: chance or common ancestry?, Science 214 (1981) 149–159.
[53] N. Nocerino, F.K. Soong, L.R. Rabiner, D.H. Klatt, Comparative study of several distortion measures for speech recognition, IEEE Proc. Int. Conf. Acoustics, Speech, and Signal Processing 11.4.1 (1985) 387–390.

About the Author—TUAN D. PHAM received his Ph.D. in civil engineering from the University of New South Wales in 1995 (Australia). He is currently an Associate Professor in the School of Information Technology and Director of Bioinformatics Applications Research Centre at James Cook University. Dr. Pham has published two research books and more than 100 refereed journal and conference papers in the fields of engineering numerical analysis (fuzzy finite element methods), soft computing, pattern recognition, image processing, speaker recognition, bioinformatics, and computational biology.