Similarity Searching in DNA Sequences by Spectral Distortion Measures

Tuan D. Pham[1,2]

[1] Bioinformatics Applications Research Centre
[2] School of Information Technology
James Cook University
Townsville, QLD 4811, Australia
tuan.pham@jcu.edu.au

Abstract. Searching for similarity among biological sequences is an important research area of bioinformatics because it can provide insight into the evolutionary and genetic relationships between species that open doors to new scientific discoveries such as drug design and treament. In this paper, we introduce a novel measure of similarity between two biological sequences without the need of alignment. The method is based on the concept of spectral distortion measures developed for signal processing. The proposed method was tested using a set of six DNA sequences taken from *Escherichia coli* K-12 and *Shigella flexneri*, and one random sequence. It was further tested with a complex dataset of 40 DNA sequences taken from the GenBank sequence database. The results obtained from the proposed method are found superior to some existing methods for similarity measure of DNA sequences.

1 Introduction

Given the importance of research into methodologies for computing similarity among biological sequences, there have been a number of computational and statistical methods for the comparison of biological sequences developed over the past decade. However, it still remains a challenging problem for the research community of computational biology [1,2]. Two distinct bioinformatic methodologies for studying the similarity/dissimilarity of sequences are known as alignment-based and alignment-free methods. The search for optimal solutions using sequence alignment-based methods is encountered with difficulty in computational aspect with regard to large biological databases. 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 [3], such as the Euclidean distance [4], Euclidean and Mahalanobis distances [5], Markov chain models and Kullback-Leibler discrepancy (KLD) [6], cosine distance [7], Kolmogorov complexity [8], and chaos theory [9]. Our previous work [10] on sequence comparison has some strong similarity to the work by Wu et al. [6], in which statistical measures

P. Perner (Ed.): ICDM 2006, LNAI 4065, pp. 24–37, 2006.

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. [6] 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.

Among alignment-free methods for computing distances between biological sequences, there seems rarely any work that directly computes distances between biological sequences using the concept of a distortion measure (error matching). If a distortion model can be constructed for two biological sequences, we can readily measure the similarity between these two sequences. In addition, based on the principles that spectral distortion measures are derived [11], their use is robust for handling signals subjected to noise and having significantly different lengths; and for extracting good features in order to enable the task of a pattern classifier much more effective.

In this paper we are interested in the novel application of some spectral distortion measures to obtain solutions to difficult problems in computational biology: i) studying the relationships between different DNA sequences for biologcal inference, and ii) searching for similar library sequences stored in a database to a given query sequence. These tasks are designed to be carried out in such a way that the computation is efficient and does not depend on sequence alignment.

In the following sections we will firstly discuss how a DNA sequence can be represented as a sequence of corresponding numerical values; secondly we will then address how we can extract the spectral feature of DNA sequences using the method of linear predictive coding; thirdly we will present the concept of distortion measures of any pair of DNA sequences, which serve as the basis for the computation of sequence similarity. 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 [6,10].

2 Numerical Representation of Biological Sequences

One of the problems that hinder the application of signal processing to biological sequence analysis is that either DNA or protein sequences are represented by characters and thus do not make themselves ready for numerical signal-processing based methods [16,17]. One available and mathematically sound model for converting a character-based biological sequence into a numeral-based biological one is the resonant recognition model (RRM) [12,13]. We therefore adopted the RRM to implement the novel application of the linear predictive coding and its cepstral distortion measures for DNA sequence analysis.

The resonant recognition model (RRM) is a physical and mathematical model which can extract protein or DNA sequences using signal analysis methods. 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 electron-ion interaction potential (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 [12,14,15]:

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

Where q is a change of momentum of the delocalised electron in the intreaction 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 (see Table 1).

Numerical series obtained this way are then analyzed by digital signal analysis methods in order to extract information adequate to the biological function. Discrete Fourier transform (DFT) is applied to convert the numerical sequence to the frequency domain sequence. After that, for the purpose of extracting mutual spectral characteristics of sequences, having the same or similar biological function, cross-spectral function is used:

$$S_n = X_n Y_n^* \qquad n = 1, 2, \ldots, \frac{N}{2} \tag{3}$$

where X_n is the DFT coefficients of the x_m, Y_n^* is the complex conjugate DFT coefficients of the $y(m)$. Based on the above cross-spectral function, we can obtain a spectrum. In the spectrum, peak frequencies, which are assumed that mutual spectral frequency of two analyzed sequences, can be observed [13].

Additionally, when we want to examine the mutual frequency components for a group of protein sequences, we usually need to calculate the absolute values of multiple cross-spectral function coefficients M:

$$|M_n| = |X1_n| \cdot |X1_n| \ldots |XM_n| \qquad n = 1, 2, \ldots, \frac{N}{2} \tag{4}$$

Furthermore, a signal-to-noise ratio (SNR) of the consensus spectrum (the multiple cross-spectral function for a large group of sequences with the same biological function, which has been named *consensus spectrum* [13]), is found as a magnitude of the largest frequency component relative to the mean value of the spectrum. The peak frequency component in the consensus spectrum is considered to be significant if the value of the SNR is at least 20 [13]. Significant frequency component is the characteristic RRM frequency for the entire group of biological sequences, having the same biological function, since it is the strongest frequency component common to all of the biological sequences from that particular functional group.

Table 1. Electron-Ion Interaction Potential (EIIP) values for nucleotides and amino acids [13,15]

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

Apart from this approach to the analysis of biological sequences, the RRM also offers some physical explanation of the selective interactions between biological macromolecules, based on their structure. The RRM considers that these selective interactions (that is the recognition of a target molecule by another molecule, for example, recognition of a promoter by RNA polymerase) are caused by resonant electromagnetic energy exchange, hence the name *resonant recognition model*. According to the RRM, the charge that is being transferred along the backbone of a macromolecule travels through the changing electric field described by a sequence of EIIPs, causing the radiation of some small amount of electromagnetic energy at particular frequencies that can be recognized by other molecules. So far, the RRM has had some success in terms of designing a new spectral analysis of biological sequences (DNA/protein sequences) [13].

3 Spectral Features of DNA Sequences

Having pointed out that the difficulty for the application of signal processing to the analysis of biological data is that it deals with numerical sequences rather than character strings. 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 sequences, we will apply the principle of linear predictive coding (LPC) to extract the spectral feature of a DNA sequence known as the LPC cepstral coefficients, which have been successfully used for speech recognition.

We are motivated to explore the use of the LPC model because, in general, time-series signals analyzed by the LPC have several advantanges as follows. First, the LPC is an analytically tractable model which is mathematically precise and simple for computer implementation. Second, the LPC model and its LPC-based distortion measures have been proved to give excellent solutions to many problems concerining with pattern recognition [19].

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 combination of the past p samples. This linear prediction can be expressed as [18,19]

$$\hat{s}(n) = \sum_{k=1}^{p} a_k \, s(n-k) \tag{5}$$

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

$$e(n) = s(n) - \hat{s}(n) = s(n) - \sum_{k=1}^{p} a_k \, s(n-k) \tag{6}$$

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

$$E = \sum_{n=1}^{N} e^2(n) = \sum_{n=1}^{N} \left[s(n) - \sum_{k=1}^{p} a_k \, s(n-k) \right]^2 \tag{7}$$

To solve (7) for the prediction coefficients, we differentiate E with respect to each a_k and equate the result to zero:

$$\frac{\partial E}{\partial a_k} = 0, \; k = 1, \ldots, p \tag{8}$$

The result is a set of p linear equations

$$\sum_{k=1}^{p} a_k \, r(|m-k|) = r(m), \; m = 1, \ldots, p \tag{9}$$

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

$$r(m) = \sum_{n=1}^{N-m} s(n)\, s(n+m),\ m = 0, \dots, p \tag{10}$$

Equation (9) can be expressed in matrix form as

$$\mathbf{R}\,\mathbf{a} = \mathbf{r} \tag{11}$$

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:

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

$$\mathbf{a}^T = \begin{bmatrix} a_1 \; a_2 \; a_3 \cdots a_p \end{bmatrix}$$

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

$$\mathbf{r}^T = \begin{bmatrix} r(1) \; r(2) \; r(3) \cdots r(p) \end{bmatrix}$$

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

Thus, the LPC coefficients can be obtained by solving

$$\mathbf{a} = \mathbf{R}^{-1}\,\mathbf{r} \tag{12}$$

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

3.2 LPC Cepstral Coefficients

If we can determine the linear prediction coefficients 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 [19].

$$c_0 = \ln(G^2) \tag{13}$$

$$c_m = a_m + \sum_{k=1}^{m-1} \left(\frac{k}{m}\right) c_k a_{m-k},\ 1 \leq m \leq p \tag{14}$$

$$c_m = \sum_{k=1}^{m-1} \left(\frac{k}{m}\right) c_k a_{m-k},\ m > p \tag{15}$$

where G is the LPC gain, whose squared term is given as [20]

$$G^2 = r(0) - \sum_{k=1}^{p} a_k r(k) \tag{16}$$

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 [19], 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: $0 \leq d(\mathbf{x}, \mathbf{y}) < \infty$, $\mathbf{x}, \mathbf{y} \in V$ 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: $d(\mathbf{x}, \mathbf{z}) \leq d(\mathbf{x}, \mathbf{y}) + d(\mathbf{y}, \mathbf{z})$ 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 [19] 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 [21]

$$\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} D(\mathbf{x}_i, \mathbf{y}_i) \tag{17}$$

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 [21]:

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

$$e(n) = -\sum_{i=0}^{p} a_i s(n-i) \tag{18}$$

where $a_0 = -1$.

The sum of squared errors can be now expressed as

$$\begin{aligned} E &= \sum_{n=0}^{N-1+p} e^2(n) + \sum_{n=0}^{N-1+p} \left[-\sum_{i=0}^{p} a_i s(n-i)\right]\left[-\sum_{j=0}^{p} a_j s(n-j)\right] \\ &= \sum_{i=0}^{p} a_i \sum_{j=0}^{p} a_j \sum_{n=0}^{N-1+p} s(n-i)s(n-j) \end{aligned} \tag{19}$$

We also have

$$\sum_{n=0}^{N-1+p} s(n-i)s(n-j) = \sum_{n=0}^{N-1+p} s(n)s(n-j+i) = r(|i-j|) \tag{20}$$

Therefore,

$$E = \sum_{i=0}^{p} a_i \sum_{j=0}^{p} a_j r(|i-j|) = \mathbf{a}^T \mathbf{R}_s \mathbf{a} \tag{21}$$

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

$$e'(n) = -\sum_{i=0}^{p} a_i s'(n-i) \tag{22}$$

where $a_0 = -1$.

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

$$E' = \sum_{i=0}^{p} a_i \sum_{j=0}^{p} a_j r'(|i-j|) = \mathbf{a}^T \mathbf{R}_{s'} \mathbf{a} \tag{23}$$

where

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

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

$$D = \frac{E'}{E} = \frac{\mathbf{a}^T \mathbf{R}_{s'} \mathbf{a}}{\mathbf{a}^T \mathbf{R}_s \mathbf{a}} \geq 1 \tag{24}$$

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 [19]

$$V(\omega) = \log S(\omega) - \log S'(\omega) \tag{25}$$

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 [22]

$$D_{IS} = \int_{-\pi}^{\pi} [e^{V(\omega)} - V(\omega) - 1] \frac{d\omega}{2\pi} = \int_{-\pi}^{\pi} \frac{S(\omega)}{S'(\omega)} \frac{d\omega}{2\pi} - \log \frac{\sigma_\infty^2}{\sigma_\infty'^2} - 1 \tag{26}$$

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

$$\sigma_\infty^2 \approx \exp\left\{ \int_{-\pi}^{\pi} \log S(\omega) \frac{d\omega}{2\pi} \right\}. \tag{27}$$

It was pointed out that the Itakura-Saito distortion measure is connected with many statistical and information theories [19] 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 [19]

$$D_{LR} = \frac{\mathbf{a}'^T \, \mathbf{R}_s \, \mathbf{a}'}{\mathbf{a}^T \, \mathbf{R}_s \, \mathbf{a}} - 1 \tag{28}$$

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:

$$\log S(\omega) = \sum_{n=-\infty}^{\infty} c_n \, e^{-jn\omega} \tag{29}$$

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 [23], the L_2-norm cepstral distance between $S(\omega)$ and $S'(\omega)$ can be related to the root-mean-square log spectral distance as [19]

$$D_c^2 = \int_{-\pi}^{\pi} |\log S(\omega) - \log S'(\omega)|^2 \frac{d\omega}{2\pi} = \sum_{n=-\infty}^{\infty} (c_n - c'_n)^2 \tag{30}$$

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 (30) can be truncated to some finite number $L \geq p$, that is

$$D_c^2(L) = \sum_{m=1}^{L} (c_m - c'_m) \tag{31}$$

5 Experiments

We have carried 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 Escherichia coli K-12 and from Shigella flexneri; and one random sequence. The second test involves a complex set of 40 DNA sequences, which was used for searching similar sequences to a query sequence.

5.1 Phylogenetic Study of DNA Sequences

The algorithm was tested with 6 DNA sequences, taken from the threonine operons of Escherichia coli K-12 (gi:1786181) and Shigella 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 3362798 (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). In addition, we compared all six sequences with a randomly generated sequence (rand-thrA), using the same length and base composition as *ec*-thrA.

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 [24]. The multiple sequence alignment has then been used to calculate an identity matrix and the distance matrix using DNADist from the PHYLIP package [25] and the modification of the Kimura distance model [26]. The DNADist program uses nucleotide sequences to compute a distance matrix, under the modified Kimura model of nucleotide substitution. Being similiar to the Jukes and Cantor model [27], which constructs the transition probability matrix based on the assumption that a base change is independent of its identity, the Kimura 2-paramter 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 [10] and the chaos game representation [9] 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 rand-thrA closer to the two thrA sequences, whose relationship is not supposed to be so.

5.2 Database Searching of Similar Sequences

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

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 (Homo sapiens 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).

Sensitivity and selectivity were computed to evaluate and compare the performance of the proposed models with other distance measures [6]. 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. [6], 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. Both sensitivity and selectivity obtained from SimMM 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.

6 Conclusions

Comparison between sequences is a key step in bioinformatics when analyzing similarities of functions and properties of different sequences. Similarly, evolutionary homology is analyzed by comparing DNA and protein sequences. So far, most such analyses are conducted by aligning first the sequences and then comparing at each position the variation or similarity of the sequences. Multiple sequence alignments of several hundred sequences is thereby always a bottleneck, first due to long computational time, and second due to possible bias of multiple sequence alignments for multiple occurrences of highly similar sequences. An alignment-free comparison method is therefore of great value as it reduces the technical constraints as only pairwise comparisons are necessary, and is free of bias. Non-alignment methods are designed to compare each pair unrelated to other pairwise comparisons, and the distortion measures can compute pair-wise sequence similarity in such fashion. Given an appropriate numerical representation of DNA sequences, the performance of the new approach for DNA sequence comparison has been found to be better than that of other existing non-alignment methods. Spectral distortion measures are computationally efficient, mathematically tractable, and physically meaningful.

Some issues for future investigations will include further exploration of models for numeral representation of biological sequences – the current experimental results analyzed by the LPC-based distortion measures are affected by the RRM which is not a unique way for expressing character-based biological sequence in terms of numerical values. The application of vector quantization (VQ) [21] of LPC coefficients, where the distance measure is the distance between two LPC vectors, can be a potential approach for improving the calculation of similarity. This can also be readily extended to the use of VQ-based hidden Markov models [19] for similarity searching.

References

1. Ewens, W.J. and Grant,G.R.: Statistical Methods in Bioinformatics. Springer, NY, 2001.
2. Miller,W.: Comparison of genomic DNA sequences: solved and unsolved problems. Bioinformatics **17** (2001) 391397.
3. Vinga,S. and Almeida,J.: Alignment-free sequence comparisona review. Bioinformatics **19** (2003) 513523.
4. Blaisdell, B.E.: Ameasure of the similarity of sets of sequences not requiring sequence alignment. Proc. Natl Acad. Sci. USA **83** (1986) 51555159.
5. Wu,T.J., Burke,J.P. and Davison,D.B.: A measure of DNA sequence dissimilarity based on Mahalanobis distance between frequencies of words. Biometrics **53** (1997) 14311439.
6. Wu,T.J., Hsieh,Y.C. and Li,L.A.: Statistical measures of DNA dissimilarity under Markov chain models of base composition. Biometrics **57** (2001) 441448.
7. Stuart,G.W., Moffett,K. and Baker,S.: Integrated gene and species phylogenies from unaligned whole genome protein sequences. Bioinformatics **18** (2002) 100108.
8. Li,M., Badger,J.H., Chen,X., Kwong,S., Kearney,P. and Zhang,H.: An information-based sequence distance and its application to whole mitochondrial genome phylogeny. Bioinformatics **17** (2001) 149154.
9. Almeida,J.S., Carrico,J.A., Maretzek,A., Noble,P.A. and Fletcher,M.: Analysis of genomic sequences by chaos game representation. Bioinformatics **17** (2001) 429437.
10. Pham, T.D., and Zuegg, J.: A probabilistic measure for alignment-free sequence comparison. Bioinformatics **20** (2004) 34553461.

11. Nocerino, N., Soong, F.K., Rabiner, L.R. and D.H. Klatt, D.H.: Comparative study of several distortion measures for speech recognition, *IEEE Proc. Int. Conf. Acoustics, Speech, and Signal Processing* **11.4.1** (1985) 387-390.
12. Veljkovic, V. and Slavic, I: General model of pseudopotentials, *Physical Review Lett.* **29** (1972) pp. 105-108.
13. Cosic, I.: Macromolecular bioactivity: Is it resonant interaction between macromolecules? – theory and applications, *IEEE trans. Biomedical Engineering* **41** (1994) 1101-1114.
14. Veljkovic, V., Cosic, I., Dimitrijevic, B. and Lalovic, D.: 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. C.H. de Trad, Q. Fang, and I. Cosic, Protein sequence comparison based on the wavelet transform approach, *Protein Engineering* **15** (2002) 193-203.
16. Anatassiou, D.: Frequency-domain analysis of biomolecular sequences, *Bioinformatics* **16** (2000) 1073-1082.
17. Anatassiou, D.: Genomic signal processing, *IEEE Signal Processing Magazine* **18** (2001) 8-20.
18. Makhoul, J.: Linear prediction: a tutorial review, *Proc. IEEE* **63** (1975) 561-580.
19. Rabiner, L. and Juang, B.H.: *Fundamentals of Speech Recognition.* New Jersey, Prentice Hall, 1993.
20. Ingle, V.K. and Proakis, J.G.: *Digital Signal Processing Using Matlab V.4.* Boston, PWS Publishing, 1997.
21. Gray, R.M.: Vector quantization, *IEEE ASSP Mag.* **1** (1984) 4-29.
22. Itakura, F. and S. Saito, S.: A statistical method for estimation of speech spectral density and formant frequencies, *Electronics and Communications in Japan* **53A** (1970) 36-43.
23. O'Shaughnessy, D.: *Speech Communication – Human and Machine.* Reading, Massachusetts, Addison-Wesley, 1987.
24. Thompson, J.D., Higgins, D.G. and Gibson, T.J.: CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. **22** (1994) 4673-4680.
25. Felsenstein, J.: PHYLIP (Phylogeny Inference Package), version 3.5c. Distributed by the Author, Department of Genetics, University of Washington, Seattle, WA, 1993.
26. Kimura, M.: A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. **16** (1980) 111-120.
27. Jukes, T.H. and Cantor, C.R.: Evolution of protein molecules. In Munro,H.N. (ed.), Mammalian Protein Metabolism. Academic Press, NY, pp. 21-132, 1969.