**MATCH**
*Communications in Mathematical and in Computer Chemistry*

MATCH Commun. Math. Comput. Chem. **65** (2011) 477-488

ISSN 0340 - 6253

# A New Measure for Similarity Searching in DNA Sequences

Yusen Zhang* and Wei Chen
*School of Mathematics and Statistics, Shandong University at Weihai*
*Weihai 264209, China*

(Received June 14, 2010)

**Abstract.**
The purpose of the present study was designed to develop a new mathematical model for comparison of DNA sequences. Instead of the classical distances, a new distance based on dinucleotide absolute frequency in large DNA sequences is introduced. The proposed distance that requires neither homologous sequences nor prior sequence alignments is used to search for similar sequences from a database. This method was tested using a set of 39 DNA sequences and a set of 63 DNA sequences. The sensitivity and the selectivity are computed to evaluate and compare the performance of the proposed distance measure. Real data analysis shows that it is a very efficient, high-selective and high-sensitive comparison algorithm that can determine the relative dissimilarity in a large dataset of DNA sequences very rapidly.

## 1 Introduction

Sequence comparison is a fundamental task in Computational Biology that aims to discover similarity relationships between molecular sequences. Searching database with a DNA sequence rely heavily on sequence comparison techniques. Because of the importance of research into similarity measure, a number of efficient algorithms have been developed for searching genetic databases for biologically significant similarities in DNA sequences. The traditional algorithms for comparing biological sequences are based mostly on the technique of sequence alignment [9, 28]. Such approaches have been hitherto widely used. Nevertheless, sequence alignment considers only local mutations of the genome, therefore it is not suitable to measure events and mutations that involve longer segments of genomic sequences. For this reason many alignment-free distance measures have been recently

*Corresponding author: zhangys@sdu.edu.cn

introduced [4–7, 14, 16, 17, 27, 29, 30]. Up to now, many efficient alignment-free measures have been proposed.

Methods for alignment-free sequence comparison of biological sequences utilize several concepts of distance measures [23], such as the Euclidean distance [2], Euclidean and Mahalanobis distances [24], Markov chain models and Kullback- Leibler discrepancy (extended KLD) [25], cosine distance [21], Kolmogorov complexity [18], Lempel-Ziv (LZ) complexity [19], chaos theory [1] and statistical measures [8, 15, 20]. The statistical measure SimMM [20] was performed using the Mahalanobis distance and the standardized Euclidean distance under Markov chain model of base composition, as well as the extended KLD [25] and SK-LD [26]. In order to evaluate them under Markov chain model of base composition, all the initial and transition probabilities need to be estimated using the whole query sequence and the Mahalanobis distance in practice may be too difficult to compute when the word sizes increase.

In this study, a new distance based on dinucleotide absolute frequency in large DNA sequences is introduced. We associate a 16-component vector with a DNA sequence. The components of the vector indicate the absolute frequency of dinucleotide. Then the comparison of DNA sequences is transformed into a simpler comparison of vectors. Instead of the classical distances, a weighted squared Euclidean distance is used to measure the distance between two vectors. The weighted function, called stabilized function, may help to promote the performance of the distance measure by adjusting parameter $m$ when desired. The weighted absolute frequency algorithm comparing with several word-based methods has better performances in sensitivity and selectivity and can help to better determine the relative dissimilarity of large dataset of genetic sequences.

# 2 Methods and algorithms

## 2.1 Absolute frequency

Consider a DNA sequence L read from the 5'- to the 3'-end with $n$ bases. By considering neighboring two bases, we can obtain sixteen dinucleotide XY: AA, AC, AG, AT, CA, CC, CG, CT, GA, GC, GG , GT, TA, TC, TG, and TT. The cumulative numbers of the nucleotide X denoted by the positive integer $F_X$ and the cumulative numbers of the dinucleotide XY denoted by the positive integer $F_{XY}$. The absolute frequency $P_L(\text{XY})$ is defined as the ratio of the cumulative numbers of the dinucleotide XY to that of the first

nucleotide X. That is

$$P_L(\text{XY}) = \frac{F_{XY}}{F_X}.$$

For DNA sequence L, the dinucleotide absolute frequency vector is defined by:

$$V_L = [P_L(\text{AA}), P_L(\text{AC}), P_L(\text{AG}), ..., P_L(\text{TT})],$$

By this way, we get a correspondence between the DNA sequence and a 16-component vector $V_L$. A DNA sequence can be analyzed by studying the corresponding dinucleotide absolute frequency vector.

## 2.2 Dissimilarity measure

Given two strands of DNA sequences $Q$ and $L$ (for the query and a library sequence in a database), let

$$V_L = [P_L(\text{AA}), P_L(\text{AC}), P_L(\text{AG}), ..., P_L(\text{TT})]$$

be the dinucleotide absolute frequency vector over a segment $W_L$, which is a window of length $l + 1$ from the sequence $L$. And set

$$V_Q = [P_Q(\text{AA}), P_Q(\text{AC}), P_Q(\text{AG}), ..., P_Q(\text{TT})],$$

where $W_Q$ be defined similarly for $Q$. Define

$$M(L) = \Sigma_{XY} P_L(\text{XY}) f_m(P_L(\text{XY})),$$

where the sum extends over all dinucleotide XY and the weighted function $f_m(x)$ is a stabilized function that is implemented to promote the performance of the distance measure by adjusting parameter $m$ .

Define the distance between the two segment $W_L$ and $W_Q$ by

$$\Delta M(L, Q) = M(L - Q) = \Sigma_{XY}(P_L(\text{XY}) - P_Q(\text{XY}))f_m(P_L(\text{XY}) - P_Q(\text{XY})).$$

Thus, $\Delta M(L, Q)$ quantities the difference between the distributions L and Q. In what follows, $W_L$ and $W_Q$ are shifted over $L$ and $Q$, respectively. A distance (say, window

distance) is taken for each pair $W = (W_L, W_Q)$. The distance between $L$ and $Q$ is taken to be the minimum of all window distances. That means the $AFd$ (absolute frequency distance) is

$$AFd(L,Q) = \min_{W}\{AFd_W(L,Q)\},$$

with

$$AFd_W(L,Q) = \Delta M(L,Q)'\Delta M(L,Q) = \Sigma_{XY}[(P_L(\text{XY}) - P_Q(\text{XY}))f_m(P_L(\text{XY}) - P_Q(\text{XY}))]^2,$$

where the sum extends over all dinucleotide XY.

For each library sequence $L$, we choose the sliding window length $l_W$ to be the minimum of the length of $L$ and the length of the query sequence $Q$. The window is shifted from left to right over the longer sequence. Let step sliding window is $u\%$, the first window starts at Position 1, the second at $\frac{u}{100}l_W + 1$, the third at $\frac{2u}{100}l_W + 1$, and so on. Hence, we have $(1 - \frac{u}{100})$ overlap on the windows.

A comparison between a pair of DNA sequences to judge their similarities and dissimilarities can be carried out by calculating the distance $AFd(L,Q)$. The analysis of similarity among each library sequence $L$ and the query sequence $Q$ is based on the assumption that the smaller is the distance $AFd(L,Q)$ the more similar are the two sequences.

The rest of work is how to select the stabilized function for different applications. The similarity search is to search a database of known function sequences and uses the structures and functions of the most closely matched known sequences to analyze the query sequence. For this application, we would use following typical stabilized functions:

$$f(x) = \frac{1}{(1+x)^m},$$

where $m$ is a nonnegative integer.

# 3 Comparing the Performances of Dissimilarity Measures

## 3.1 Evaluation methods

Sensitivity and selectivity were computed to evaluate and compare the performance of the proposed distance measure $AFd$ with other distance measures in previous studies [20, 25].

All mentioned distance measures were used to perform a search for similarities of the query sequence HSLIPAS $Q$ (of length 1612) human lipoprotein lipase (LPL) against a test dataset of $s$ sequences ($t$ HSLIPAS-related sequences and $s - t$ HSLIPAS-unrelated sequences).

Sensitivity is expressed by the number of HSLIPAS related sequences found among the first closest $t$ 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.

## 3.2 Experiment no.1

In order to compare the performance of several word-based methods used in [24] with our proposed method, the proposed distance $AFd$ was used to search for similar sequences of a query sequence from a complex dataset 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. This dataset has been studied in [20,25]. These 39 sequences were selected from mammals, viruses, plants, etc., of which lengths vary between 322 and 14121 bases.

HSLIPAS is also used as the query sequence. 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,

13790 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, 14121 bp) and HUMAPOBF (human apolipoprotein B-100 mRNA, complete cds, 10089 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).

Before computing the sensitivity and selectivity by using our proposed method, a series of steps of sliding window $5\% - 30\%$ performed for our approach, and the results listed in Table 1. It shows that this approach that appear to produce the high and steady sensitivity and sensitivity vales when steps of sliding window among $10\% - 30\%$. Whereas both sensitivity and selectivity obtained from our proposed method were of 18 sequences. These results agree with those obtained using the KLD of Markov models [20] and better than those obtained by using the recommended standardized Euclidean distance under the Markov chain models of base composition, of which sensitivity and selectivity were of 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

Table 1: Comparison of sensitivity and selectivity for dataset of 39 library sequences

| m | step (u %) | 5 % | 8 % | 10 % | 12 % | 14 % | 16 % | 18 % | 20 % | 22 % | 24 % | 26 % | 28 % | 30 % |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Sensitivity | 19 | 18 | 19 | 19 | 19 | 19 | 19 | 19 | 19 | 19 | 19 | 19 | 19 |
| | Selectivity | 18 | 14 | 18 | 18 | 18 | 17 | 18 | 16 | 18 | 18 | 17 | 16 | 17 |
| 1 | Sensitivity | 19 | 18 | 19 | 19 | 19 | 19 | 19 | 19 | 19 | 19 | 19 | 19 | 19 |
| | Selectivity | 18 | 16 | 18 | 18 | 18 | 17 | 18 | 18 | 17 | 18 | 18 | 18 | 18 |
| 2 | Sensitivity | 19 | 19 | 19 | 19 | 19 | 19 | 19 | 19 | 19 | 19 | 19 | 19 | 19 |
| | Selectivity | 18 | 17 | 18 | 18 | 18 | 18 | 18 | 17 | 17 | 18 | 18 | 18 | 18 |
| 3 | Sensitivity | 19 | 19 | 19 | 19 | 18 | 19 | 19 | 19 | 19 | 19 | 19 | 19 | 19 |
| | Selectivity | 18 | 17 | 18 | 18 | 17 | 18 | 18 | 17 | 17 | 18 | 18 | 18 | 18 |
| 4 | Sensitivity | 19 | 19 | 19 | 19 | 18 | 19 | 19 | 19 | 19 | 19 | 18 | 19 | 19 |
| | Selectivity | 18 | 17 | 18 | 17 | 17 | 17 | 18 | 17 | 17 | 17 | 18 | 18 | 18 |
| 5,6 | Sensitivity | 19 | 18 | 19 | 19 | 18 | 19 | 19 | 19 | 19 | 19 | 19 | 19 | 19 |
| | Selectivity | 18 | 18 | 18 | 16 | 17 | 16 | 18 | 18 | 17 | 16 | 18 | 18 | 18 |
| 7 | Sensitivity | 19 | 18 | 19 | 19 | 18 | 19 | 19 | 19 | 19 | 19 | 19 | 19 | 19 |
| | Selectivity | 17 | 18 | 18 | 16 | 17 | 16 | 18 | 19 | 17 | 16 | 18 | 18 | 18 |
| 8,9 | Sensitivity | 19 | 18 | 19 | 19 | 18 | 19 | 19 | 19 | 19 | 19 | 19 | 19 | 19 |
| | Selectivity | 17 | 18 | 18 | 16 | 17 | 16 | 18 | 19 | 17 | 16 | 18 | 19 | 18 |
| 10,11, | Sensitivity | 19 | 18 | 19 | 19 | 19 | 19 | 19 | 19 | 19 | 19 | 19 | 19 | 19 |
| 12 | Selectivity | 17 | 18 | 18 | 16 | 18 | 16 | 18 | 19 | 18 | 16 | 18 | 19 | 18 |
| 13 | Sensitivity | 19 | 18 | 19 | 19 | 19 | 19 | 19 | 19 | 19 | 19 | 19 | 19 | 19 |
| | Selectivity | 17 | 18 | 18 | 16 | 18 | 16 | 19 | 19 | 18 | 16 | 18 | 19 | 19 |
| 14,16, | Sensitivity | 19 | 18 | 19 | 19 | 19 | 19 | 19 | 19 | 19 | 19 | 19 | 19 | 19 |
| 19 | Selectivity | 17 | 18 | 19 | 16 | 18 | 17 | 19 | 19 | 18 | 16 | 18 | 19 | 19 |
| 26 | Sensitivity | 19 | 19 | 19 | 19 | 19 | 19 | 19 | 19 | 19 | 19 | 19 | 18 | 19 |
| | Selectivity | 19 | 18 | 19 | 17 | 18 | 17 | 19 | 19 | 19 | 16 | 19 | 18 | 19 |
| 83 | Sensitivity | 19 | 19 | 19 | 19 | 19 | 19 | 19 | 19 | 19 | 19 | 19 | 19 | 19 |
| | Selectivity | 18 | 19 | 19 | 17 | 18 | 18 | 19 | 19 | 19 | 19 | 18 | 19 | 19 |
| 95 | Sensitivity | 19 | 18 | 19 | 19 | 19 | 19 | 19 | 19 | 19 | 19 | 19 | 19 | 19 |
| | Selectivity | 18 | 19 | 19 | 17 | 17 | 18 | 19 | 19 | 18 | 19 | 18 | 19 | 19 |
| 99 | Sensitivity | 19 | 19 | 19 | 19 | 19 | 19 | 19 | 19 | 19 | 19 | 19 | 19 | 19 |
| | Selectivity | 18 | 19 | 19 | 17 | 17 | 18 | 19 | 19 | 18 | 19 | 18 | 19 | 18 |
| 999 | Sensitivity | 19 | 19 | 19 | 19 | 19 | 19 | 19 | 19 | 19 | 19 | 19 | 19 | 19 |
| | Selectivity | 19 | 19 | 19 | 16 | 16 | 19 | 19 | 19 | 16 | 17 | 17 | 19 | 18 |

different word sizes were combined [25]. The false rejections given by the proposed method are similar, all are HSBMHSP, whereas the false acceptances are AA2CG.

### 3.3 Experiment no.2

The proposed distance $AFd$ was further tested with a more complex dataset of 63 DNA sequences taken from the GenBank sequence database. These 63 sequences were selected from mammals, invertebrates, viruses, plants, bacteria, etc., of which lengths vary between 322 to 2 462 499 bases. Every member of the test dataset is classified as being related or not related in biological function to the query sequence. There are 35 sequences classified as being related, and 28 sequences classified as being not related.

Wu et al. [26] use both SK-LD and BLAST to perform a search for dissimilarities/similarities of the query sequence HSLIPAS (1612 bp) human lipoprotein lipase against this test dataset. The SK-LD and BLAST scores between HSLIPAS and 63 library sequences are sorted from the highest to lowest similarity, respectively, and the sensitivity and selectivity are used to quantify their performances.

They obtained that the sensitivity and selectivity for SK-LD are 34 and 30, respectively, and those for BLAST are 29 and 22, respectively, at the default parameter setting and are no better than 33 and 28, respectively, at other parameter settings (the optimal result is obtained). Hence, SK-LD performs better than BLAST. Also, SK-LD improves the combined K-LD [25], whose sensitivity and selectivity are 31 and 24, respectively. They also computed the sensitivity and selectivity of SimMM of Pham and Zuegg [20] that are 32 and 26, respectively.

Tables 2 show the results of our approach. Observing Table 2, we find that all the sensitivity is 33 sequences and all the selectivity is no less than 28 sequences when $m > 0$ and steps of sliding window is set between $10\% - 30\%$ (that means $70\% - 90\%$ overlap on the windows). As it can be easily observed, the best sensitivity and selectivity are of 33 and 33 sequences, respectively. Two similar false rejections are HSBMHSP (13793 bp Human sapiens mRNA for basement membrane heparan) and PTLPL2 (1018 bp Pan troglodytes (chimpanzee) lipoprotein lipase gene, exon 6); whereas one similar false acceptances are AA2CG (4675 bp Adeno-associated virus 2, complete genome), the others are one or two ANANIFBH (5936 bp Anabaena PCC7120 nitrogenase, ferrodoxin-like protein nifS, nifU, and nitrogenase reductase genes, complete cds), AL3HP (1786 bp Bacteriophage alpha-3

Table 2: Comparison of sensitivity and selectivity dataset of 63 library sequences

| m | step (u %) | 5 % | 8 % | 10 % | 12 % | 14 % | 16 % | 18 % | 20 % | 22 % | 24 % | 26 % | 28 % | 30 % |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Sensitivity | 33 | 32 | 33 | 33 | 33 | 33 | 33 | 33 | 33 | 32 | 33 | 33 | 33 |
| | Selectivity | 29 | 25 | 30 | 29 | 28 | 29 | 27 | 29 | 28 | 27 | 28 | 27 | 27 |
| 1 | Sensitivity | 33 | 32 | 33 | 33 | 33 | 33 | 33 | 33 | 33 | 33 | 33 | 33 | 33 |
| | Selectivity | 29 | 28 | 30 | 29 | 30 | 28 | 30 | 29 | 30 | 28 | 30 | 30 | 29 |
| 2 | Sensitivity | 33 | 33 | 33 | 33 | 33 | 33 | 33 | 33 | 33 | 33 | 33 | 33 | 33 |
| | Selectivity | 29 | 30 | 30 | 29 | 31 | 29 | 30 | 30 | 29 | 28 | 31 | 31 | 30 |
| 10 | Sensitivity | 33 | 33 | 33 | 33 | 33 | 33 | 33 | 33 | 33 | 33 | 33 | 33 | 33 |
| | Selectivity | 31 | 32 | 32 | 29 | 32 | 28 | 32 | 33 | 30 | 28 | 32 | 33 | 31 |
| 14 | Sensitivity | 33 | 33 | 33 | 33 | 33 | 33 | 33 | 33 | 33 | 33 | 33 | 33 | 33 |
| | Selectivity | 31 | 32 | 33 | 29 | 32 | 30 | 33 | 33 | 31 | 28 | 32 | 33 | 32 |
| 83 | Sensitivity | 33 | 33 | 33 | 33 | 33 | 33 | 33 | 33 | 33 | 33 | 33 | 33 | 33 |
| | Selectivity | 30 | 31 | 30 | 30 | 31 | 28 | 32 | 28 | 29 | 30 | 32 | 33 | 32 |

H protein gene, complete cds) or AGGGLINE (7360 bp A.geoffroyi gamma-globin gene and L1 LINE element).

The prediction accuracy will generally increase in the beginning but will not increase all the way when the value of $m$ increases. From Table 2 we can see that the best predicted accuracies have been gotten when the value of $m$ is 14.

# 4 Conclusions

we advocate the use of dinucleotide absolute frequency within a DNA sequence and stabilized function as a basis for structuring the distance measure $AFd$. The adoption of appropriate stabilized function can greatly improve the prediction accuracy. Here we use $\frac{1}{(1+x)^m}$ as stabilized function. In fact, we can also use other function as stabilized function, for example, when we replace $\frac{1}{(1+x)^m}$ with $(1-x)^m$ in this application, the result is near consistent. It is also noteworthy that the use of dinucleotide absolute frequency is less sensitive to length than that of dinucleotide frequency of DNA sequence.

we have demonstrated experimentally the ability of $AFd$ to detect biologically significant matches between a query and large datasets of DNA sequences while varying stabilized function. The comparison demonstrates that $AFd$ is a simple, high-sensitive, and high-selective method of rapid sequence comparison that can detect novel sequence relationships. This can significantly enhance the current technology in comparing large datasets of DNA sequences.

## Acknowledgements

The authors thank Prof. Tuan D. Pham for providing all MATLAB code for SimMM and Prof. Tiee-Jian Wu for his kindly help. The authors also thank Dr. Qi Dai for his technical help. This work was supported in part by the Shandong Natural Science Foundation (Y2006A14).

## References

[1] 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.

[2] B. E. Blaisdell, A measure of the similarity of sets of sequences not requiring sequence alignment, *Proc. Natl Acad. Sci.* **83** (1986) 5155–5159.

[3] B. E. Blaisdell, Effectiveness of measures requiring and not requiring prior sequence alignment for estimating the dissimilarity of natural sequences, *J. Mol. Evol.* **29** (1989) 526–537.

[4] D. Bielinska-Waz, P. Waz, T. Clark, Similarity studies of DNA sequences using genetic methods, *Chem. Phys. Lett.* **445** (2007) 68–73.

[5] C. Burge, A. M. Campbell, S. Karlin, Over– and under–representation of short oligonucleotides in DNA sequences, *Proc. Natl. Acad. Sci.* **89** (1992) 1358–1362.

[6] W. Chen, Y. Zhang, Comparisons of DNA sequences based on dinucleotide, *MATCH Commun. Math. Comput. Chem.* **61** (2009) 533–540.

[7] W. Chen, Y. Zhang, Three distances for rapid similarity analysis of DNA sequences, *MATCH Commun. Math. Comput. Chem.* **61** (2009) 781–788.

[8] Q. Dai, Y. Yang, T. Wang, Markov model plus k-word distributions: A synergy that produces novel statistical measures for sequence comparison, *Bioinformatics* **24** (2008) 2296–2302.

[9] D. Davison, Sequence similarity searching for molecular biologists, *Bull. Math. Biol.* **46** (1984) 437–474.

[10] O. Gotoh, Y. Tagashira, Locations of frequently opening regions on natural DNAs and their relation to functional loci, *Biopolymers* **20** (1981) 1033–1042.

[11] W. A. Hide, L. Chan, W. H. Li, A review of the structural, functional, and evolutionary relationships of the lipase gene superfamily, *J. Lipid Res.* **33** (1992) 167–178.

[12] W. Hide, J. Burke, D. Davison, Biological evaluation of d2, an algorithm for high performance sequence comparison, *J. Comput. Biol.* **1** (1994) 199–215.

[13] S. C. Johnson, Hierarchical clustering schemes, *Psychometrika* **2** (1967) 241–254.

[14] S. Karlin, I. Ladunga, Comparisions of eukaryotic genomic sequences, *Proc. Natl. Acad. Sci.* **91** (1994) 12832–12836.

[15] M. R. Kantorovitz, G. E. Robinson, S. Sinha, A statistical method for alignment–free comparison of regulatory sequences, *Bioinformatics* **23** (2007) 1249–1255.

[16] B. Liao, K. Ding, A graphical approach to analyzing DNA sequences, *J. Comput. Chem.* **26** (2005) 1519–1523.

[17] Z. Liu , B. Liao, W. Zhu, A new method to analyze the similarity based on dual nucleotides of the DNA sequence, *MATCH Commun. Math. Comput. Chem.* **61** (2009) 541–552.

[18] 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.

[19] H. H. Otu, K. Sayood, A new sequence distance measure for phylogenetic tree construction, *Bioinformatics* **19** (2003) 2122–2130.

[20] T. D. Pham, J. Zuegg, A probabilistic measure for alignment–free sequence comparison, *Chem. Phys. Lett.* **20** (2004) 3455–3461.

[21] G. W. Stuart, K. Moffett, S. Baker, Integrated gene and species phylogenies from unaligned whole genome protein sequences, *Bioinformatics* **18** (2002) 100–108.

[22] O. C. Uhlenbeck, P. N. Borer, B. Dengler, I. J. Tinoco, Stability of RNA hairpin loops: A 6 -C m -U 6, *J. Mol. Biol.* **73** (1973) 483–496.

[23] S. Vinga, J. Almeida, Alignment–free sequence comparison – A review, *Bioinformatics* **19** (2003) 513–523.

[24] 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.

[25] 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.

[26] T. J. Wu, Y. C. Hsieh, L. A. Li, Dissimilarity measures and estimation of the degree of dissimilarity between DNA sequences, *Bioinformatics* **21** (2005) 4125–4132.

[27] R. Wu, R. Li, B. Liao, G. Yue, A novel method for visualizing and analyzing DNA sequences, *MATCH Commun. Math. Comput. Chem.* **63** (2010) 679–690.

[28] M. S. Waterman, *Mathematical Methods for DNA Sequences*, CRC Press, Boca Raton, 1989.

[29] J. F. Yu, J. H. Wang, X. Sun, Analysis of similarities/dissimilarities of DNA sequences based on a novel graphical representation, *MATCH Commun. Math. Comput. Chem.* **63** (2010) 493–512.

[30] Y. Zhang, A simple method to construct the similarity matrices of DNA sequences, *MATCH Commun. Math. Comput. Chem.* **60** (2008) 313–324.