Proposed Better Sequence Alignment for Identification of Organisms Using DNA Barcode

Sandeep Kaur, Sukhamrit Kaur and Sandeep K. Sood

Abstract DNA barcoding is a system that uses short sequence instead of whole genome; hence, it makes ecological system more accessible. It provides fast and accurate species identification. DNA barcoding, for many applications such as natural resource preservation, water quality monitoring, disease vector identification, generates short DNA sequence from standard region of genome known as marker. Methods such as BLAST, FASTA, and Smith–Waterman are generally employed for species identification using DNA barcoding. Among these methods, BLAST is used for fast species identification but gives less accurate results as compared to Smith–Waterman which is a very slow process. BLAST has been performed using a sequence to study the effect of word size on accuracy. Its results show more accuracy with smaller word size. In this study, a new algorithm with combined features of both BLAST and Smith–Waterman methods has been purposed and implemented which include more numbers of hits/matches. These hits vary with word size and threshold.

Keywords BLAST · CO1 · DNA · DNA barcode · FASTA · Identification Marker

Abbreviation

A	Adenine
BOLD	Barcode of life data
BLAST	Basic Local Search Alignment Tool
CBOL	Consortium for barcode of life
CO1	Cytochrome c oxidase 1
CPU	Central processing unit

S. Kaur (✉) · S.K. Sood
Department of Computer Science, Guru Nanak Dev University Regional Campus, Gurdaspur, Punjab, India
e-mail: sandeep.gndu18@gmail.com

S. Kaur
Department of Computer Science, Shanti Devi Arya Mahila College Dinanagar, Gurdaspur, Punjab, India

B. Panda et al. (eds.), *Innovations in Computational Intelligence*, Studies in Computational Intelligence 713, https://doi.org/10.1007/978-981-10-4555-4_9

DNA	Deoxyribonucleic acid
FASTA	Fast alignment
FISH-BOL	Fish Barcode of Life
G	Guanine
Ibol	International barcode of life
ITS	Internal transcribed spacer
matK	Megakaryocyte-associated tyrosine kinase
mtDNA	Mitochondrial deoxyribonucleic acid
MUSCLE	Multiple sequence comparison by log-expectation
rRNA	Ribosomal ribonucleic acid
T	Thymine

1 Introduction to Bioinformatics

In word bioinformatics, bio means molecular and informatics is related to information technology. It means applying the informatics techniques to analyze and interpret the information related to the molecules. It is a management information system for molecular biology and can be used in medical science to identify the gene causing the disease [1]. Today, biology is playing a vital role and will be important in the coming years. Biology when depends on chemistry becomes biochemistry. When there is a need to explain biological process at atomic level, it becomes biophysics.

There is huge data collected by biologists which needs to be analyzed and interpreted using some tools of computer science resulting in new field: bioinformatics. Large storage is needed for producing large amount of data at fast rate [2, 3]. For example, growth of sequence data of GenBank is shown in Fig. 1.

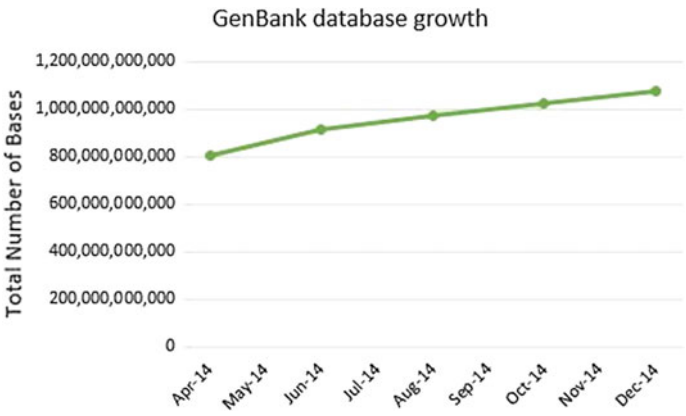

Fig. 1 Number of bases from April to December 2014 [34]

This leads to great need of information system for easy management of large amount of data. The experimental laboratory is producing over 100 gigabytes of data every day. With this large amount of data produced, improvements have been done in CPU, disk storage, and Internet for faster computations. The aims of bioinformatics are organizing the data to make it easily accessible for researchers and developing the tools for analysis and interpretation of that data [2].

Sources of data associated with the molecules can be as follows: DNA sequences made of four base letters (A, C, T, and G) of 1000 bases long, protein sequences of size 300 amino acid; macromolecular structure of size 1000 atomic coordinates; genomes of 3 billion bases; and gene expressions. The computational techniques in bioinformatics include sequence alignment, database design and data mining, phylogenetic tree (evolutionary tree) construction, functions and predicting the structure of protein, gene finding, and expression data clustering [2].

1.1 Applications of Bioinformatics

Bioinformatics has many applications in the real world in various fields such as basic research areas, medicine, microbiology, and also in agriculture. Bioinformatics helps in studying the genome comparing genomes.

Basic Research Areas

Basic research areas include functional genomics (analyzing the integration, coordination, and functions of all genes present in organism), evolutionary genomics (done by comparing the genomes of species to see how two species are relating to each other), proteomics (study of proteins), systems biology, and high-performance computing [4].

Medicine

Bioinformatics is helpful in the field of medicine such as in drug discovery, personalized medicine, and preventive medicine [4]. In drug discovery, identification of specific drugs for particular purpose is done. Personalized medicine involves the study of gene behavior of individual due to some specific medicine, i.e., how they interact with medicine. Preventive medicine involves preventing disease rather than curing it, i.e., curing the disease at early or initial stages.

Microbiology

It is the study of genomes of microorganisms to understand their behavior or environment, energy, health, and other industrial applications.

Agriculture

Bioinformatics can help in agriculture by producing better and stronger crops that can bear drought, disease by sequencing the genomes of plants [4].

2 Introduction to DNA Barcoding

Monitoring the biological effects of global climate, identification of organisms has become important to preserve species because of increasing habitat destruction. We know very less about diversity of plants and animals that are living on earth. There is an estimation of 5–50 million plants and animal species out of which less than 2 million have been identified. Yearly rate of extinction has increased from one species per million to 100–1000 species per million which means thousands of plants and animals are lost each year, most of which are not identified yet [5]. The high levels of destruction and endangerment of ecosystem have lead to improved system for identifying species. In recent years, a new ecological (preservation) approach called DNA barcoding has been proposed to identify species and ecology research [4, 6]. DNA barcoding is a system for fast and accurate species identification which will make ecological system more accessible [7]. DNA barcoding is a best tool for those who are not experts but they can do it easily using this tool even for those that are very different from each other and difficult to identify [8].

DNA barcoding first came to attention of the scientific community in 2003 when Paul Hebert's science research group at University of Guelph published a paper titled "biological identifications through DNA barcodes." DNA barcoding is a tool for identification of species and for taxonomic research. DNA barcoding is not a new concept as Carl Woese used rRNA and molecular markers such as rDNA and mtDNA to discover archaea, i.e., prokaryotes, and then for drawing evolutionary tree. But DNA barcoding uses short DNA sequence instead of whole genome for eukaryotes. Species can be identified using a short section of DNA from standard region of genome to generate DNA barcode. DNA barcode is short DNA sequence made of four nucleotide bases A (adenine), T (thymine), C (cytosine), and G (guanine). Each base is represented by a unique color in DNA barcode as shown in Fig. 1. Even nonexperts can identify species from small, damaged, or industrially processed material [9] (Fig. 2).

Identification of small damaged or processed material or dietary supplements can be done using DNA barcoding because it is a DNA-based method that uses DNA for its processing and identification, and as we know that degradation does not affect DNA of organism as it affects protein. So, because of this reason, the DNA

Fig. 2 DNA barcode [35]

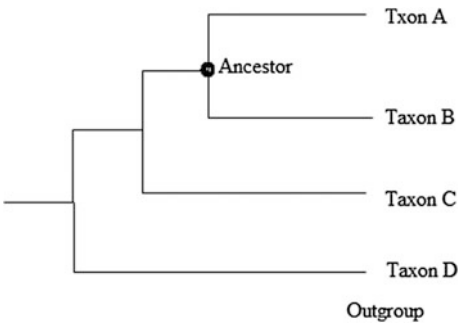

Fig. 3 Evolutionary tree

barcoding can be done using any sample even from egg to adult that means in any life stage of organisms [10].

DNA barcode should be generated from individual section of DNA. This standard or individual section also known as marker varies among the species. In animals, Paul Hebert proposed the use of CO1 or cox1 present in mitochondrial gene as marker for generating barcode, and now, it is recognized by International Barcode of Life (IBOL) as official marker for animals. The main reason for choosing mitochondrial gene is because of its small intraspecific and large interspecific differences. But CO1 is not suitable for other group of organisms because it is uniform in them. So Internal Transcribed Spacer (ITS) is recognized for fungus, and two genes from chloroplast genome, rbcl and matK, are recognized as barcode markers for plants by IBOL [6, 11].

The sequence data generated by sequencer are used for identification and to construct a phylogenetic tree, in which related individuals are clustered together. Phylogenetic tree or evolutionary tree is a branching diagram which represents the evolutionary history of species. It can provide large amount of information. For a particular species, tree can identify ancestors and closest relatives of species. With the help of evolutionary trees many questions can be answered such as what kind of earliest animals look like or which features are inherited by their descendants [12, 13] (Fig. 3).

2.1 Applications of DNA Barcoding

1. *Controlling agricultural pests*

Pest damage in agriculture can cost farmers billion dollars. DNA barcoding can help with this problem by identifying pests in any stage of life which makes it easier to control them. The global Tephritid barcoding initiative contributes to management of fruit flies by providing tools to identify and stop fruit flies at border.

2. *Identifying disease vectors*

Vector species causes many serious animal and human infectious diseases like malaria. DNA barcoding allows nonecologists to identify these vector species to understand these diseases and cure them. A global mosquito barcoding initiative in building a reference barcode library can help public health officials to control these diseases causing vector species more effectively with very less use of insecticides.

3. *Sustaining natural resources*

Overharvesting of natural resources such as hardwood trees and fishes causes species extinction and economy collapse of industries that rely on them. Using DNA barcoding, natural resource managers can monitor illegal trade of products that made of these natural resources. The FISH-BOL reference barcode is a library for hardwood trees, to improve the management and conservation of natural resources.

4. *Protecting endangered species*

Primate population is reduced by 90% in Africa because of bushmeat hunting. DNA barcoding can be used by law enforcement to bushmeat in local markets which is obtained from bushmeat.

5. *Monitoring water quality*

Drinking water is a process resource for living being. By studying organism living in lakes, rivers, and streams, their health can be measured or determined. DNA barcoding is used to create a library of these species that can be difficult to identify. Barcoding can be used by environmental agencies to improve determination of quality and to create better policies which can ensure safe supply of drinking water.

6. *Routine authentication of natural health products*

Authenticity of natural health products is an important, economic, health, and conservation issue. Natural health products are often considered as safe because of their natural origin.

7. *Identification of plant leaves even if flowers or fruit are not available.*
8. *Identification of medicinal plants* [14].

2.2 Procedure of DNA Barcoding

DNA barcoding mainly have two purposes:

(a) Building the barcode library of identified species.
(b) Matching the barcode sequence of the unknown sample with the barcode library for its identification.

First of all, a specimen is collected from organism for generating the data known as DNA barcode. This specimen can be either preserved in museum or can be collected from live field. The sample then goes through laboratory processes which involve tissue sampling, DNA extraction, and sequencing. In tissue sampling, the tissue is collected from the specimen, and then, DNA is extracted from the tissue. The extracted DNA is then sequenced using sequencer. Before the sequencing of DNA, the genes are amplified polymerase chain reaction (PCR). After sequencing, a DNA sequence is generated from extracted DNA, known as DNA barcode [15]. Generally, DNA barcode is approximately 300–1000 bases in length. Here, we are not taking length in base pairs because the DNA barcode is generated in the form of only one strand of DNA, so we have considered it only in bases instead of base pairs. DNA barcode is visually represented by chromatogram. The DNA barcode can be used for two purposes: First, DNA barcode can be stored in any database for future references and to build a barcode library. For this purpose, ecologic expertise is required for selecting one or several individuals per species as reference samples in the barcode library. Second purpose is that it can be analyzed for information stored in it. This information can be used to identify the specimen for which DNA barcode is used as query sequence. This query sequence is compared with other nucleotide sequences that already exist in databases such as GenBank, BOLD. The process of comparing the sequences is known as sequence alignment. Various sequence alignment algorithms are used for sequence alignment such as pairwise sequence alignment algorithms and multiple sequence alignment. In my work, I have focused on pairwise sequence alignment and some of its important algorithms [15] (Figs. 4 and 5).

2.3 *Users of DNA Barcoding*

As we know that DNA barcoding is a tool for identifying and classifying species and to create an online database containing DNA sequences of various organisms that can be used as future reference. If DNA sequence extracted from the sample is matched with any sequence of database, then sample is identified; otherwise, it is a

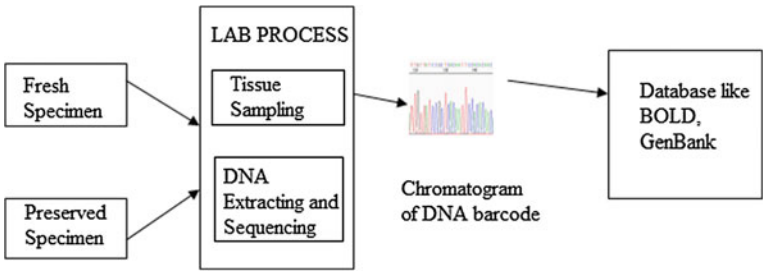

Fig. 4 DNA barcoding procedure

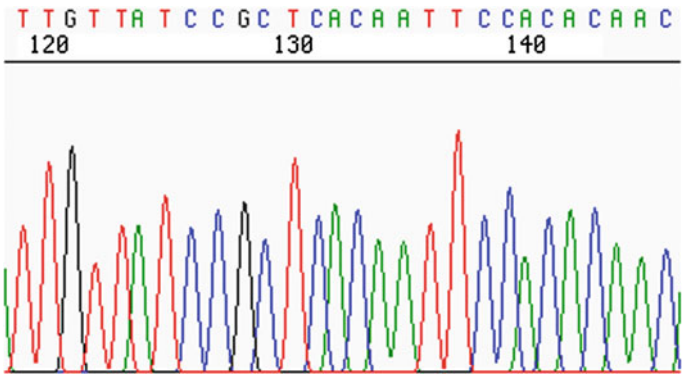

Fig. 5 Chromatogram of DNA barcode generated by sequencer [36]

new species that is not yet discovered or identified. DNA barcoding is a tool that is very helpful for users for different purposes.

2.3.1 Taxonomist

Taxonomist is a biologist that studies the different organisms and categorizes them according to the relationship between them. So, DNA barcoding is very helpful for taxonomists as finding the relationship becomes very easy with it. A taxonomist can use DNA barcoding for studying the nature and its distributions for which he or she must have good skills and can gather information from literature, museum, or databases [16].

2.3.2 Ecologists

Ecologists study the environment, organisms living in that environment, and how actions of human affect other living organisms. DNA barcoding is a great tool to identify the nature and its distribution. Information needed for this purpose can be gathered from literatures, museums, databases.

2.3.3 Conservationist

Conservationist is a person who is a member of conservation system to conserve the environment. Their work is to study the environment and identify the species living in the environment. Information needed for identification of organisms is gathered from the images, databases, taxonomists which is also a user of DNA barcoding [16].

2.3.4 Legal Users

Legal users are like police. These users need identification for the purpose of forensic investigation, controlling the crimes related to wildlife, and illegal trades. Information needed for this purpose can be gathered from images, databases, and taxonomists.

2.3.5 Animal/Human Health

The users that deal with animal and human health also use the benefits of DNA barcoding. These are the persons who find the medicinal and harmful properties in the organisms because some animals can be very helpful for other organisms and in making good medicines. The sources of information needed for this purpose are images, databases, and taxonomists. Their skills and interest are medium, i.e., varying [16].

2.3.6 Environmental Protection

The user that deals with environment protection is the persons that do the identification of invasive species and indicator species. Invasive species are those which are dangerous to other organisms in environment and have harmful effects on them, e.g., pests that have harmful and dangerous effects on the species on which they are living. For example, fungus and bacteria are pests. The indicator species are those species whose presence in an area indicates some environmental effects on other species in the same environment. Like, due to some changes or conditions occurring in that area, scientists can examine the indicator species, to check that how these changes or conditions have affected the other species that are very difficult to examine. The information needed for this purpose can be gathered from the images, databases, and taxonomists. Their skills and interest are not good and not bad.

2.3.7 Naturalist

Naturalists are the users that deal with the environment. They identify the organisms living in the environment to study the environment and its classification and must have good skills. The information needed for this purpose can be collected from literature, databases, and taxonomists.

2.3.8 Users that Deal with Utilization of Biodiversity

These users deal with the identification of plants, crops, or animals for the purpose of their utilization, so that they can be utilized by other organisms in the

environment. The information needed for this purpose can be gathered from images, field guides, taxonomists, and online databases.

2.3.9 Public

These users are people from public. Some of them may be interested and other may be uninterested. They do identification due to some occasional purpose like some people want to study for their thesis or other research work. The source of information needed for identification is images or guides [16].

2.4 Future of DNA Barcoding

DNA barcoding is like a revolution in history of identification of organisms. DNA-based identification has linked the various samples collected from the field or museum with the sequences present in the online databases such as BOLD (Barcode of Life Data) and GenBank. Various questions have asked by many biologists about its validity and its benefits in the future. As long as there is a need for conservation of nature and natural resources and its right utilization, there is a need of identification of organisms [16]. DNA barcoding is a much better identification tool than the traditional ones and previous methods. DNA barcoding has been succeeded for distinguishing and identifying the organisms living in the environment. Also it has been helpful in case of cryptic species which are closely related to each other or looks same but slight biological difference. DNA barcoding is fast and more accurate tool which can be used to build a device to identify the species based upon the input sample with in few minutes or seconds [17].

2.5 Limitation in DNA Barcoding

The aim of DNA barcoding is identification of samples and species discovery. Identification of specimen involves the process of naming the unidentified specimen using DNA barcoding. There are many benefits of DNA barcoding but have some limitations too. [18] Accuracy of diagnosis of an organism is dependent on the intraspecific variation which is compared with intraspecific differences of other species. With this using a short DNA sequence, we can check that in which taxon a particular species or organism fits. This decision should be made with confidence [19]. Some limitations of DNA barcoding are given below.

2.5.1 Unclear Objective Hypothesis

The main aim of DNA barcoding is to build a DNA barcode reference library, then validating and testing the library for future use and last exploring the unidentified organisms. Validation and testing of reference library and exploring the unidentified species are mixed because the reference library should be tested and must be corrected and validated. The identification of organisms depends on the data of reference library. So, if the reference library is not tested, then it will affect the identification.

2.5.2 Incorrect Identification of Samples

The second limitation is the human error and inaccuracies made while creating the reference library. These errors can become big problems so reference library should be tested. With this, there can be conflicts between identification of same species by different persons. Like when two different persons are working in different laboratories but on same taxa, then the errors made by them in reference library can result in different identification for same taxa. So, there is a need for correct maintenance of reference library data.

2.5.3 Interpretation of Species Identification

The term species identification causes a big confusion in species discovery and sample identification. The actual meaning of species identification is identification of sample to species. To avoid or minimize the confusion, the concept of species discovery and sample identification should be clearly defined [18].

3 Literature Survey

Luscombe [2] hase explained the need of computer resources in bioinformatics. It has been explained that large amount of data and information is produced which needs to be handled, managed, analyzed, and interpreted. This can be done by using tools available in bioinformatics. Bioinformatics employs number of computational techniques such as sequence alignment, data mining, evolutionary tree generating, genome sequences, gene findings, and prediction of protein structure.

Hebert et al. [20] have stated that identification of organism lies in the construction of systems that uses DNA sequences as barcodes. In this paper, author has established that mitochondrial gene cytochrome c oxidase 1 (CO1) can serve as

main part of identification system for animals. CO1 has taken from mitochondrial gene of genome in animals and has 100% successful in identification of animals. CO1 identification system will provide a reliable, accessible, and cost-effective solution to the current problem of species identification.

Cohen [3] has explored the importance of computer science in bioinformatics. In this paper, author has introduced computer scientists with the new field bioinformatics. Bioinformatics has born with the needs of biologists for collecting, managing, analyzing, and interpreting data. This all can be done with resources available in computer science such as hard disk, CPU, and Internet.

Cohen [3] explained the importance of information technology in biology or medical field. Nowadays, there are lot of experiments going on in many laboratories in the world such as exploring the genes, their function, medicine field and their effect on genes, study of human genome, and genomes of other organisms. These experiments are carried out by biologists. In these experiments, huge amount of data is produced which needs to be handled, needs to be stored somewhere, and needs to be processed. For these purposes, there is a need of information technology and its resources such as storage, CPU, computing power in biology. Biologists develop some algorithms for operations to be done using information technology resources. Use of information technology in biology is known as bioinformatics. Bioinformatics reduces the complexity of these experiments and operations which are to be done in these experiments. The author has described the software developed for biologists and currently being used by them and some areas of computer science that is very important in field of bioinformatics.

Hebert and Gregory [7] have discussed that DNA barcoding is an appropriate system developed to provide fast, accurate, and automatable species identification by using short DNA sequence generated from standardized region of gene. It makes taxonomic system more accessible, beneficial to ecologists and conservationists. One day, DNA barcoding will lead us to a state when everyone will have easy access to name and biological attributes of any species on the planet. It also highlights diversity in species and may represent new species. Even though it is beneficial, it also has been controversial in some scientific areas.

Notredame [21] has discussed about multiple sequence alignment and its various algorithms. These algorithms are based on different scoring schemes such as matrix-based scoring schemes that include algorithms like ClustalW and consistency-based scoring schemes that include algorithms like T-Coffee. Then, the procedures of these algorithms are also discussed.

Little et al. [22] have concluded that to use DNA sequences for species identification, an algorithm to compare the sequences is needed. Two novel alignment-free algorithms were used to identify query sequences for the purpose of DNA barcoding. Gymnosperm nrITS2 and plastid matK sequences were used on test data. Results show that DNA barcoding could be used to identify samples with a very less error. Geographic range can be used as elimination factor without which DNA barcoding does not appear to be useful for species-level identification.

Waugh [23] discussed that since 250 years lot of work in systematic has done but majority of species are still unidentified. Increase in extinction rates and need of biological monitoring lead to DNA barcoding. So DNA barcoding is a technology that has been proposed that might expedite the species identification. This method involves various markers for various species for efficient results. For this, there is a need of various identification programmers to be coordinate. DNA barcoding may prove to be very useful taxonomic tool.

Vaidya et al. [1] have stated that cancer is widely spreading worldwide. One of the types of cancer is breast cancer. In this paper, bioinformatics and its process are explained. It includes that study of breast cancer involves predicting its causes, drugs for its cure, and predicting its transmission in the next generations.

Dalton et al. [24] have discussed that smuggling of wildlife animals for commercial purpose has lead to population decline in South Africa. So, mitochondrial CO1 gene was sequenced to determine species of unknown sample in three suspects of South African forensic wildlife cases. Two unknown samples were identified as domestic cattle and third was identified as common reedbuck.

Kress and Erickson [15] have stated that DNA barcoding is a new method for fast identification of species used in DNA based on DNA sequences, generated from small standardized region of genome. As a research tool for ecologists, it expands ability to diagnose species by including all life-history stages of organisms. The DNA barcoding involves building of DNA barcode library of known species and then matching barcode sequence of unknown sample against the barcode library for identification. It has grown because number of sequences has generated as barcode and in terms of its application.

Nagy et al. [25] have concluded that DNA barcoding is now a popular and well-accepted tool for identification of various species and detection of taxonomic diversity, since its introduction in 2003. This method is becoming an essential part of taxonomic practice. DNA barcoding is a tool that makes species identification easier and also bodies such as iBOL (Barcode of Life Project) and CBOL (Consortium for the Barcode of life) for their projects.

Abilash and Rohitaktha [26] conclude that pairwise sequence alignment is one of the methods to arrange two biological sequences to identify similarity which indicates functional, structural relationship between them. Pairwise alignment has two methods: local and global alignment methods. Local alignment is applicable in searching local similarity in large sequences. Global alignment aligns the sequences by taking whole sequence at a time. Local and global pairwise alignment methods are analyzed to find out similarity between the sequences.

Khallaf et al. [27] stated that due to substitution of species, accurate identification of seafood species in markets is a growing concern. It has become prime priority of governments to identify the already processed fish products. DNA barcoding was applied to some samples purchased from Egyptian markets and were analyzed. Sequencing of mitochondrial cytochrome c oxidase (CO1) gene revealed 33.3% species substitution in fish products which demonstrates that DNA barcoding is a reliable tool for detecting fish products.

4 Problem Definition

1. Alignment of DNA sequences for checking similarity using small word size and gaps.
2. Comparison of various sequence alignment algorithms.

4.1 Sequence Alignment

Sequence alignment is a process of comparing two or more sequences, whether DNA, RNA, or protein sequence, to look for similar patterns in sequences [13, 28]. DNA sequence is made of four bases A (adenine), T (thymine), C (cytosine), and G (guanine), and for identification of species, these need to be aligned. Comparison of sequences has become very helpful in understanding the information content and functions of genetic sequence and can tell that how much the sequences are closely related. Sequence alignment provides solution to many problems in bioinformatics including identification of new species, finding relationship between species, and for predicting the function and structure of genes and proteins [29]. Sequence alignment is a process of looking for similarity regions in two sequences. Aligning two sequences is done to get a most suitable alignment. Suitable alignment means aligning the two sequences in such a way that it gives maximum score to do alignment. The alignment is done to look for similarity in two sequences. If there are more matches in sequences, then sequences are more similar, and if less matches, then less similar. While aligning the sequences, there are three possibilities: gap or indel (means insertion deletion), match, and mismatch.

Let us take an example of very small sequences:

$$\begin{aligned} \text{S1} &= \text{ATCG} \\ \text{S2} &= \text{TCA} \end{aligned}$$

We can align these sequences in various ways; with each alignment, we will get different scores, but we will consider only that which alignment gives us maximum score value. Let us suppose the values have gap, match, and mismatch. The value of match must be positive and value of gap and mismatch must be negative.

$$\begin{aligned} &\text{Let value of Match} = 2 \\ &\text{Mismatch} = -1 \\ &\text{Gap} = -2. \end{aligned}$$

First way to align is as following:

S1	A	T	C	G
S2		T	C	A
	Gap	match	match	mismatch
	-2	+2	+2	-1

Score= -2+2+2-1 = 1

Second way to align is as following:

S1	A	T	C	G
S2	T	C	A	
	Mismatch	mismatch	mismatch	Gap
	-1	-1	-1	-2

Score= -1-1-1-2 = -5

Third way to align is as following:

S1	A	T	C	G	
S2	T		C	A	
	Mismatch	gap	match	mismatch	
	-1	-2	+1	-1	S

Score= -1-2+1-1 = -3

From the above three ways, first method gives us maximum score, hence is a suitable alignment. The alignment can be done only of those sequences which are having some similarity [28].

DNA sequence alignment is of two types:

(a) Pairwise sequence alignment and
(b) Multiple sequence alignment.

(i) *Pairwise Sequence Alignment*

Pairwise sequence alignment is a method of aligning two sequences at a time. The sequences can be of same or different size or they can be somewhat similar or dissimilar sequences.

(ii) *Multiple Sequence Alignment*

Multiple sequence alignment is a process of aligning or comparing more than two or three sequences with database sequences. This is done to create evolutionary tree or phylogenetic tree and to predict the structures like predicting the structure of proteins. With multiple sequence alignment algorithms, the study and analysis of relationship between various taxa using phylogenetic or evolutionary tree has become easy. The accurate multiple sequence alignment is not possible but some heuristic algorithms are used for aligning the multiple sequence. There are two scoring techniques in multiple sequence alignment and that is matrix-based scoring scheme and consistency-based scoring scheme. The multiple sequence alignment algorithms with matrix-based scoring scheme are ClustalW, MUSCLE, and Kalign, and algorithms with consistency-based scoring scheme is T-Coffee [21].

4.2 Already Existing Pairwise Alignment Algorithms

Pairwise alignment is a process of aligning two sequences at one time to check for similarity between them. These methods are used to find the best matching local or

global alignments of two sequences. For example, if two sequences are taken from different organisms and from a common ancestor, then because of similarity, they will get aligned. The purpose of this arrangement is to determine the relationship between the biological sequences [26]. It is based on a score which is evaluated from the number of same characters in two sequences, number and length of gaps, required to align sequence so that the two sequences get aligned [30]. Alignments can be of two types: local alignment and global alignment. Global alignment technique involves the attempt to align every character in every sequence. In this, number of characters in sequences or size should be same. This approach would be time-consuming and inconvenient for longer sequences. In global alignment, the two sequences of equal length are aligned completely or globally. Full sequence is aligned to look for similarity. Local alignments are appropriate for dissimilar sequences which may contain similar character sequence. In local alignment, the two sequences that can be of equal or different length and of similar or dissimilar are aligned to look for similarity in local regions. It means two sequences are compared, and whereever similarity occurs, that local region is counted as having similarity [26].

4.2.1 Needleman–Wunsch Algorithm

The Needleman–Wunsch algorithm was published in 1970. It performs a global alignment on two nucleotide or protein sequences. This algorithm provides a method of finding the ideal global alignment of two sequences by maximizing the matches and minimizing the number of gaps that are necessary to align the two sequences. The alignment with the highest score must be the best alignment for which score matrix has to be prepared. Algorithm is as follows.

A and *B* are sequences and *Ai* and *Bj* represent the bases of sequence at position *i* and *j*.

Step 1: Score matrix is created.
Step 2: Trace backing is done.
Step 3: Compute an alignment that actually gives this score, start from the bottom right cell, and compare the value with the three possible sources (diagonal, up, and bottom) to see which it came from. If diagonal, then *Ai* and *Bj* are aligned, if up, then *Ai* is aligned with a gap, and if left, then *Bj* is aligned with a gap.

To analyze the time complexity of the Needleman–Wunsch algorithm, we have to analyze each part of the algorithm. For start filling the score matrix, time complexity of $O(M + N)$. The next step is filling other cells of the matrix with all the scores. For each cell of the matrix, three neighboring cells must be compared, which needs constant time. So, to fill the entire matrix, the time complexity is the number of entries, i.e., $O(MN)$. Finally, the trace back requires a number of steps. We can move a maximum of N rows and M columns, and thus, the complexity of this is $O(M + N)$. The second step is finding the final path which involves jumping

from cells of matching characters. Since this step can include a maximum of N cells, this step is $O(N)$.

Thus, the overall time complexity of this algorithm is

$$O(M+N)+O(MN)+O(M+N)+O(N)=O(MN)$$

Since this algorithm fills the cells of single matrix of size MN, i.e., $O(MN)$, and stores at most N positions for the trace back, i.e., $O(N)$, the total space complexity of this algorithm is $O(MN)+O(N)=O(MN)$.

(i) *Methodology*

Needleman–Wunsch algorithm is performed using two nucleotide sequences of size 368 base pairs. These sequences are collected from NCBI database of nucleotides.

Two sequences are taken as inputs. Source of query sequence is Homo sapiens and the sequence is as below:

ACAAGATGCCATTGTCCCCCGGCCTCCTGCTGCTGCTGCTCTCCGGG
GCCACGGCCACCGCTGCCCTGCCCCTGGAGGGTGGCCCCACCGGCCGA
GACAGCGAGCATATGCAGGAAGCGGCAGGAATAAGGAAAAGCAGCCT
CCTGACTTTCCTCGCTTGGTGGTTTGAGTGGACCTCCCAGGCCAGTGCC
GGGCCCCTCATAGGAGAGGAAGCTCGGGAGGTGGCCAGGCGGCAGGA
AGGCGCACCCCCCCAGCAATCCGCGCGCCGGGACAGAATGCCCTGCA
GGAACTTCTTCTGGAAGACCTTCTCCTCCTGCAAATAAAACCTCACCC
ATGAATGCTCACGCAAGTTTAATTACAGACCTGAA [31].

Source of other sequence is Tasmanian Mountain Shrimp, and the sequence is as below:

TCTTTAGATTTTATTTTTGGAGCTTGGTCTGGCATAGTAGGCACCGCC
CTAAGACTTATTATTCGGGCTGAATTAGGACAACCTGGTAGACTTATTG
TGATGATCAAATTTACAACGTGGTCGTAACAGCTCATGCTTTTGTGATA
TTTTTTTTATAGTTATGCCCATTATAATTGGTGGATTTGGAAATTGACTT
TTCCCTTAATATTAGGTGCTCCTGATATAGCTTTTCCTCGTATAAATAAT
ATAAGATTTTGACTTCTTCCACCTTCTTTAACTCTTCTCCTATCCAGAGG
AATAGTTGAAAGAGGTGTTGGCACAGGATGAACTGTTTATCCTCCTTTA
GCTGCTGGAATCGCCCATGC [31] (Fig. 6).

Needleman–Wunsch algorithm has been performed on Web site: http://www.ebi.ac.uk/Tools/psa/. This Web site provides the pairwise sequence alignment algorithms such as Needleman–Wunsch or Smith–Waterman algorithm. The score of this alignment is 197.5.

4.2.2 Smith–Waterman Algorithm

The Smith–Waterman algorithm is a well-known algorithm used for local sequence alignment, i.e., for determining similar characters or patterns between two

```
AB000263          1 -----ACAAGATGCCATTGTCCCCCGGCCTCCTGCTGCTGCTGCTCTCCG     45
                    ||||...|||.|            ||..|||  ||.|||
EMBOSS_001        1 TCTTT---AGATTTTATTTT------------TGGAGCT--TGGTCT---     30

AB000263         46 GGGCCACG--GCCACCGCTGCCCTGCCCCTGGAGGGTGGCCCCACCGGCC     93
                     |||...|  |.||||   |||||
EMBOSS_001        31 -GGCATAGTAGGCACC---GCCCT-------------------------     50

AB000263         94 GAGACAGCGAGCATATGCAGGAAGCGGCAGGAATAAGGAAAAGCAGCCTC    143
                    .|||       |.|||    .|..|||...||||.||||.||       |
EMBOSS_001       51 AAGA-------CTTAT----TATTCGGGCTGAATTAGGACAA-------C     82

AB000263        144 CT----GACTTTCCTCGCTTGGTGGT-------TTGA----GTGGACCTC    178
                    ||    |||||      .||||||.|       ||.|    ||||.|.|.
EMBOSS_001       83 CTGGTAGACTT------ATTGGTGATGATCAAATTTACAACGTGGTCGTA    126

AB000263        179 CCAGGCCAGTGC-----------------------CGGGCCCCTCATAG    204
                    .|||..|| |||                       ...||||.|.|||
EMBOSS_001      127 ACAGCTCA-TGCTTTTGTGATAATTTTTTTTATAGTTATGCCCATTATA-    174

AB000263        205 GAGAGGAAGCTCGGGA--GGTGGCCAGGCGGCA---GGAAGGCGCACCCC    249
                              |  |||||         |   |||
EMBOSS_001      175 ---------------ATTGGTGG---------ATTTGGA-----------    189

AB000263        250 CCCAGCAATCCGCGCGCCGGGACAGAATG--CCCT-----GCAGGAACTT    292
                          |||          .|||   .||  ||||     ..|||   |.
EMBOSS_001      190 ------AAT----------TGAC---TTGTTCCCTTAATATTAGG---TG    217

AB000263        293 CTTCTGGAAGACCTTCTCCTCCTGCAAATAAAA-----------------    325
                    ||.|||..|.|.|||.|||||.|..||||||.|
EMBOSS_001      218 CTCCTGATATAGCTTTTCCTCGTATAAATAATATAAGATTTTGACTTCTT    267

AB000263        326 ---------------CCTCACCCAT------GAATGCTCACGCAAGTTTA    354
                                  |.||.||.||      ||||         ||||.|
EMBOSS_001      268 CCACCTTCTTTAACTCTTCTCCTATCCAGAGGAAT---------AGTTGA    308

AB000263        355 A-------TT---ACA-GACCTGAA------------------------    368
                    |       ||   ||| ||  ||||
EMBOSS_001      309 AAGAGGTGTTGGCACAGGA--TGAACTGTTTATCCTCCTTTAGCTGCTGG    356

AB000263        369 ------------    368

EMBOSS_001      357 AATCGCCCATGC    368
```

Fig. 6 Output of Needleman–Wunsch algorithm [32]

nucleotide or protein sequences. Instead of looking at the total sequence, the Smith–Waterman algorithm compares segments of all possible lengths and checks for similarity.

The Smith–Waterman algorithm was published in 1981 and is very similar to the Needleman–Wunsch algorithm. But still it is different because it is a local sequence alignment algorithm. Instead of aligning the entire length of two sequences, it finds the local region of highest similarity.

Step 1: Score matrix is created. All cells have values either 0 or 1.
Step 2: Trace backing is done. It starts with the maximum value in score matrix.
Step 3: Now compute the alignment, the local alignment value takes the maximum value of all the three values taken in the global alignment with the value "0". And trace back starts with the maximum value in the score matrix and traverse diagonally aligning every character of both the sequences until it encounters the value "0" in the score matrix [26].

To analyze the time complexity of the Needleman–Wunsch algorithm, we have to analyze each part of the algorithm. For start filling the score matrix, time complexity of $O(M + N)$. The next step is filling other cells of the matrix with all the scores. For each cell of the matrix, three neighboring cells must be compared, which need constant time. So, to fill the entire matrix, the time complexity is the number of entries, i.e., $O(MN)$. The time complexity for the trace back is $O(MN)$. The time complexity of Smith–Waterman algorithm is same as Needleman–Wunsch algorithm.

Therefore, the total time complexity of the Smith–Waterman algorithm is

$$O(M+N)+O(MN)+O(MN)=O(MN)$$

The space complexity of the Smith–Waterman algorithm is also same as the complexity of Needleman–Wunsch algorithm. This is because same matrix is used and same amount of space for trace back is needed.

(i) *Methodology*

Smith–Waterman algorithm is performed using two nucleotide sequences of size 368 base pairs. These sequences are collected from nucleotide database of NCBI. In this algorithm, same nucleotide sequences are used that are used in Needleman–Wunsch algorithm.

Smith–Waterman algorithm has been performed on Web site: http://www.ebi.ac.uk/Tools/psa/. This Web site provides the pairwise sequence alignment algorithms such as Needleman–Wunsch or Smith–Waterman algorithm. The score of this alignment is 209.0 (Fig. 7).

4.2.3 FASTA

FASTA stands for fast alignment. FASTA is a fast searching algorithm used for comparing query sequence with database. It comes under dynamic programming and it was developed by Lipman and Pearson in 1985. FASTA is faster than Smith–Waterman and Needleman–Wunsch algorithms which are good for two-sequence comparison, but when to compare with entire database, they are very slow than FASTA. The algorithm is as follows:

```
AB000263        73 TGGAGGGTGGCC-----------CCACCGGCCGAGACAGCGAG-CATATG    110
                   |||||..|||.|           .|||||.||.|       || |.|||
EMBOSS_001      18 TGGAGCTTGGTCTGGCATAGTAGGCACCGCCCTA-------AGACTTAT-     59

AB000263       111 CAGGAAGCGGCAGGAATAAGGAAAAGCAGCCTCCT-GACTTTCCTCGCTT    159
                     .|..|||...||||.||||    ||.|||..| |||||      .||
EMBOSS_001      60 ---TATTCGGGCTGAATTAGGA----CAACCTGGTAGACTT------ATT     96

AB000263       160 GGTGGT-------TTGA----GTGGACCTCCCAGGCCAGTGCCGGGCCCC    198
                   ||||.|       ||.|    ||||.|.|..|||               |
EMBOSS_001      97 GGTGATGATCAAATTTACAACGTGGTCGTAACAG--------------C    131

AB000263       199 TCATAGGAGAGGAAGCTCGGGAG------------------GTGGCCA--    228
                   ||||          |||...|.|                  .||.|||
EMBOSS_001     132 TCAT----------GCTTTTGTGATAATTTTTTTTATAGTTATGCCCATT    171

AB000263       229 ------GGCGGCA---GGAAGGCGCACCCCCCCAGCAATCCGCGCGCCGG    269
                         ||.|| |   |||                 |||         .
EMBOSS_001     172 ATAATTGGTGG-ATTTGGA----------------AAT----------T    193

AB000263       270 GACAGAATG--CCCT-----GCAGGAACTTCTTCTGGAAGACCTTCTCCT    312
                   |||   .||  ||||     ..|||   |.||.|||..|.|.|||.||||
EMBOSS_001     194 GAC---TTGTTCCCTTAATATTAGG---TGCTCCTGATATAGCTTTTCCT    237

AB000263       313 CCTGCAAATAAAA--------------------------------CCTCA    330
                   |.|..||||||.|                                |.||.
EMBOSS_001     238 CGTATAAATAATATAAGATTTTGACTTCTTCCACCTTCTTTAACTCTTCT    287

AB000263       331 CCCAT------GAATGCTCACGCAAGTTTAA-------TT---ACA----    360
                   ||.||      ||||         ||||.||       ||   |||
EMBOSS_001     288 CCTATCCAGAGGAAT---------AGTTGAAAGAGGTGTTGGCACAGGAT    328

AB000263       361 GACCTG    366
                   ||.|||
EMBOSS_001     329 GAACTG    334
```

Fig. 7 Output of Smith–Waterman algorithm [32]

I is query sequence and *J* is test sequence.

Step 1: Identify common *k* words or simply words between *I* and *J* using a dot plot matrix. For DNA $k = 6$, i.e., 6 nucleotides.
Step 2: Score diagonals with *k* word matches, identify 10 best diagonals.
Step 3: Rescore initial region with a substitution score matrix.
Step 4: Join initial regions for gaps.
Step 5: Perform dynamic programming for final alignment.

The complexity of the FASTA algorithm depends on the size of the *k-tuples*, which means the larger the *k-tuples*, the faster the algorithm. The true complexity is not easily determined because the speed of the alignment of two sequences depends on the total number of marked cells' variable diagonals. The space complexity of

this algorithm is also *O(MN)* like the Needleman–Wunsch and Smith–Waterman because it uses a matrix. But it uses less space because not all cells in the matrix are marked.

4.2.4 BLAST

BLAST stands for Basic Local Alignment Search Tool. TBLAST algorithm was developed by Altschul, Gish, Miller, Myers, and Lipman in 1990 to increase the speed of FASTA by finding fewer and better spots of denser matching during the algorithm. BLAST concentrates on finding local regions of high similarity in alignments without gaps.

Algorithm:

Step 1: Using word search method, sequence is filtered to remove complexity regions.
Step 2: Identification of exact word match method searches the database for neighborhood word. Words having equal or greater scores than neighborhood score threshold are taken for alignment.
Step 3: Using maximum segment pair alignment method, it extends the possible match as ungapped alignment in both directions that stops at maximum score.

The complexity of the BLAST algorithm is *O(MN)*. This is the same time complexity as all of the other algorithms, but BLAST significantly reduces the numbers of segments which need to be extended so the algorithm runs faster than all the previous algorithms. Using BLAST for nucleotide sequences, DNA barcoding has been used as a tool for identification of three species in forensic wildlife in South Africa [24] and also it has revealed high level of mislabeling in fish fillets purchased from Egyptian markets [27].

(i) *Methodology*

BLAST has been performed using two nucleotide sequences of size 368 base pairs. These sequences are collected from nucleotide database of NCBI. In this algorithm, same nucleotide sequences are used that are used in Needleman–Wunsch algorithm and Smith–Waterman algorithm (Fig. 8).

BLAST has been performed on Web site: https://blast.ncbi.nlm.nih.gov/. There are numbers of alignments in the output and each alignment has different score (Table 1).

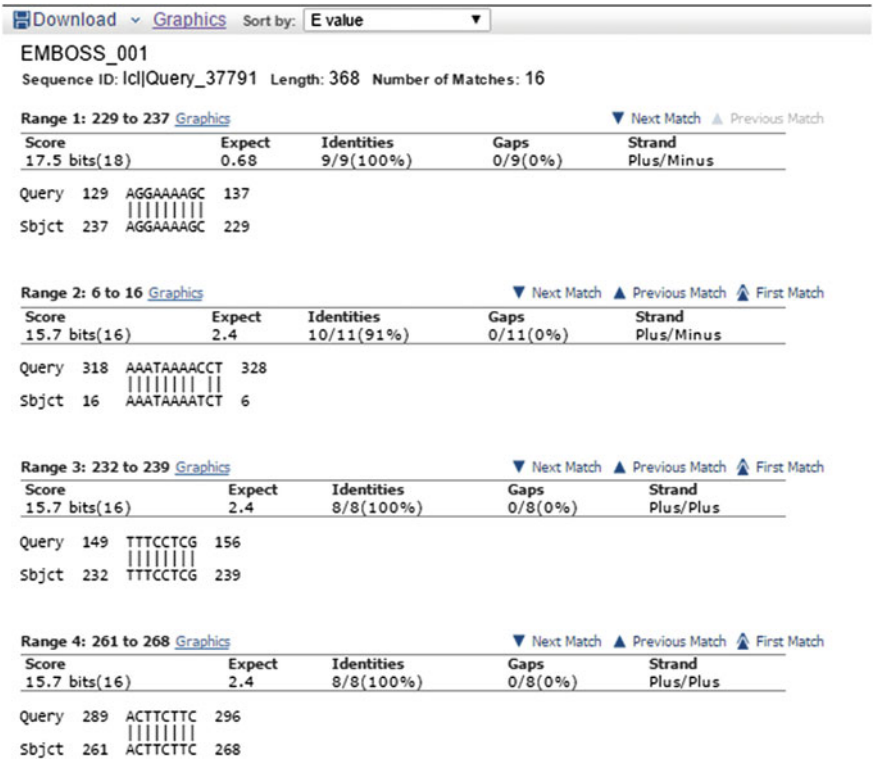

Fig. 8 Output of BLAST [33]

Table 1 Comparison of sequence alignment algorithms [26]

	Complexity	Alignment	Accuracy	Speed
Needleman–Wunsch	$O(MN)$	Global alignment	Less accurate than Smith–Waterman	Slow for searching entire database
Smith–Waterman	$O(MN)$	Local alignment	More accurate	Slow for searching entire database
FASTA	Time complexity depends on k	Local alignment	Less accurate than Smith–Waterman	Faster than above
BLAST	$O(MN)$	Local alignment	Less than Smith–Waterman	Fastest

5 Proposed Work

Proposed work of thesis is to do sequence alignment with more accuracy and this idea has been implemented with combining the features of Smith–Waterman and Basic Local Alignment Search Tool (BLAST).

In the proposed model of sequence alignment algorithm, the concept of gapped alignment from Smith–Waterman algorithm is combined with the word size and heuristic approach of BLAST algorithm. In this model, first of all, query sequence is broken into words of size 3, 4, or 5. The small size of words is to get more numbers of hits while matching because with small word the small matches cannot be missed. Then these words are stored in indexed table. Suppose we have query sequence ACTGACTGCCCGTAAATGCATCGTAGC. Now with word size *006B* = 3, underlined words are stored in the table with their indices as shown below.

ACTGACTGCCCGTAAATGCATCGTAGC

Then from the indexed table, the words are matched with sequences present in the database. Then, these words are matched with query database and aligned with insertions and deletions. Then, theses aligned words are extended to both left and right directions till the score is increasing. Finally, the highest scored pairs are chosen (Fig. 9).

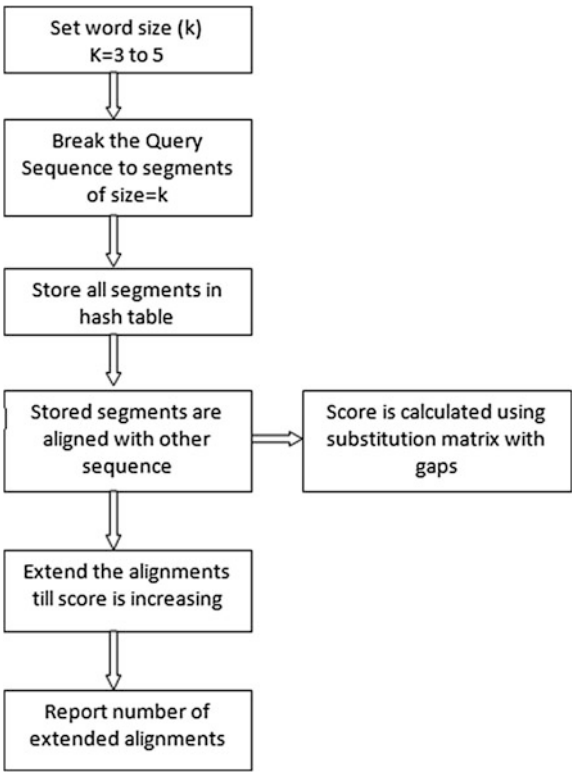

Fig. 9 Proposed work

5.1 Proposed Algorithm

Step 1: Decompose the query sequence into words of length k, use $k = 3$ to 5.
Step 2: Store all words in hash table for faster searching and matching.
Step 3: For each word, look into hash table with a score greater than threshold. The scores are calculated using a substitution matrix by including gaps in sequences. These gaps are also known as indel (insertions and deletions).
While comparing sequences A and B, if gap is inserted in B then it is known as deletion and in sequence A, at corresponding base it will be insertion.
Step 4: Search the database for sequences containing any one of the words.
Step 5: Extend the hit (matched word) in both directions until its score is increasing.
Step 6: Report the highest scoring pairs if its score is greater than the threshold.

5.2 Features in Proposed Algorithm

Proposed algorithm is the combination of best features of Smith–Waterman and BLAST algorithm. That means accuracy of identification is provided by Smith–Waterman and fast search is provided by heuristic technique of BLAST algorithm. So, the proposed algorithm provides faster search and accurate results. Also the word size used will be 3 to 5 for faster search and more sensitivity (accuracy). Because speed is directly proportional to word size and sensitivity is inversely proportional to word size large word size will give faster search speed and less sensitivity, and small word size will give less search speed and more sensitivity. So, the word size has chosen to be 3–5.

5.3 Parameters

(a) *Word Size*

Word size is the size of word taken from sequence that is used for searching in the databases. Its value to be used in the proposed algorithm will be 3–5.

(b) *Threshold*

All the words must have scored at least equal to threshold.

5.4 Methodology

For implementing the proposed work, first we have to check whether sensitivity increases with reduction in word size or not. For this, BLAST has been performed with different word size and then number of hits or number of matches is recorded. Parameters to be considered are word length or word size and number of hits in the sequence alignment. Word size denoted by k is the length of word or segment that is to be used as size of segment of sequence before starting the alignment. Number of hits is number of matches or alignments found in the sequences. To study the effect of change in work length on accuracy of species identification, BLASTN was performed using nucleotide sequence of invertebrate animal species named as Anaspides tasmaniae against nonredundant database. It returns top 100 sequences having some similarity, for each query sequence [1]. It is a freshwater species, i.e., common resident of lakes, streams, and pools in caves, in Tasmania highlands. To observe the effect of word length parameter, values of 7, 11, and 15 are used with the expect value, $E = 10$.

The sequence used for the observation which is extracted from CO1 region of Anaspides species of size 657 bases, and in FASTA format is as follows. This sequence is collected from NCBI database of nucleotides.

>EMBOSS_001

TCTTTAGATTTTATTTTTGGAGCTTGGTCTGGCATAGTAGGCACCGCC
CTAAGACTTATTATTCGGGCTGAATTAGGACAACCTGGTAGACTTATTG
GTGATGATCAAATTTACAACGTGGTCGTAACAGCTCATGCTTTTGTGAT
AATTTTTTTTATAGTTATGCCCATTATAATTGGTGGATTTGGAAATTGAC
TTGTTCCCTTAATATTAGGTGCTCCTGATATAGCTTTTCCTCGTATAAAT
AATATAAGATTTTGACTTCTTCCACCTTCTTTAACTCTTCTCCTATCCAG
AGGAATAGTTGAAAGAGGTGTTGGCACAGGATGAACTGTTTATCCTCC
TTTAGCTGCTGGAATCGCCCATGCAGGCGCTTCTGTGGACTTAGGAATT
TTTTCTCTTCATATAGCGGGAGCTTCTTCTATTTTAGGGGCGGTAAATTT
TATTACTACTTCTATTAATATGCGTGCCAATGGTATAACTTTAGATCGA
ATACCTTTATTTGTCTGATCCGTTTTTATTACTGCTATTCTTTTACTACTC
TCTCTTCCCGTTTTAGCAGGGGCAATCACAATACTTCTCACTGACCGTA
ACTTAAATACTTCTTTCTTTGACCCCGCTGGAGGAGGAGATCCATTCTT
TATCAACATAAATGCC [32] (Table 2).

Table 2 Varying number of hits with different word size

Word size (k)	No. of hits
7	518,310,295
11	32,757,086
15	14,769,504

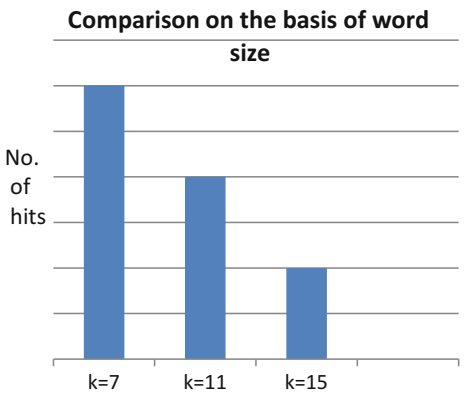

Fig. 10 Comparison on the basis of word size

The results from word size $k = 7$ returned 518,310,295 hits, $k = 11$ returned 32,757,086, i.e., less hits than returned by $k = 7$ and then $k = 15$ returned 14,769,504 which is least of all. So the observations tell us that decreasing the word size gives more number of hits, i.e., more alignments or matches, and increasing the word size gives less number of hits [33].

The implementation of proposed algorithm is done in Java. In this chapter, we will create code for proposed sequence alignment algorithm which we will run to do alignment of two sequences. The steps for creating program are (Fig. 10):

1. Define classes,
2. Define methods,
3. Write the Java code, and
4. Run and analyze the output.

The program involves data input from users, processing of that input, and producing the results. The input is taken from files containing the DNA sequences. Sequences are extracted from the files and then processed according to the program; that is, the query sequence (to be matches) is broken into segments. Size of segment is decided with choosing the word size. Word size is the size of segment in which the query sequence has to be broken. Then, these segments are matched with another sequence (with which to be matched) and alignment is done to find suitable alignments. Then, these alignments are extended to get higher scores. And then, the number of alignments which get higher score will be number of hits that are the output of this program.

5.4.1 Inputs

Two sequences are taken as inputs. These sequences are collected from NCBI database of nucleotides. Source of query sequence is Homo sapiens, and the sequence is as below:

ACAAGATGCCATTGTCCCCCGGCCTCCTGCTGCTGCTGCTCTCCGGG
GCCACGGCCACCGCTGCCCTGCCCCTGGAGGGTGGCCCCACCGGCCGA
GACAGCGAGCATATGCAGGAAGCGGCAGGAATAAGGAAAAGCAGCCT
CCTGACTTTCCTCGCTTGGTGGTTTGAGTGGACCTCCCAGGCCAGTGCC
GGGCCCCTCATAGGAGAGGAAGCTCGGGAGGTGGCCAGGCGGCAGG
AAGGCGCACCCCCCCAGCAATCCGCGCGCCGGGACAGAATGCCCTGCA
GGAACTTCTTCTGGAAGACCTTCTCCTCCTGCAAATAAAACCTCACCCA
TGAATGCTCACGCAAGTTTAATTACAGACCTGAA [31].

Source of other sequence is Tasmanian Mountain Shrimp, and the sequence is as below:

TCTTTAGATTTTATTTTTGGAGCTTGGTCTGGCATAGTAGGCACCGCC
CTAAGACTTATTATTCGGGCTGAATTAGGACAACCTGGTAGACTTATT
GTGATGATCAAATTTACAACGTGGTCGTAACAGCTCATGCTTTTGTGAT
ATTTTTTTTATAGTTATGCCCATTATAATTGGTGGATTTGGAAATTGACT
TTTCCCTTAATATTAGGTGCTCCTGATATAGCTTTTCCTCGTATAAATAA
TATAAGATTTTGACTTCTTCCACCTTCTTTAACTCTTCTCCTATCCAGA
GGAATAGTTGAAAGAGGTGTTGGCACAGGATGAACTGTTTATCCTCCTT
TAGCTGCTGGAATCGCCCATGC [31].

5.4.2 Java Packages

The packages that are used in the code are java.io and java.util. Java.io package means java input/output package which is used for taking input and producing output after processing input. All the operations related to input and output are controlled by some classes. And these classes are contained in this package. Java. util package has been used for scanner class (for taking input from user), hash table class (to create a hash table and storing segments in it), and enumeration (for retrieving the content of hash table).

5.4.3 Java Classes

The main class used for this program is ProposedAlgo which consists of all the methods defined in the program.

5.4.4 Java Methods

The program is developed in form of methods.

BuildMatrix()

This method creates a scoring matrix.

$$\text{double diagScore} = \text{score}[i-1][j-1] + \text{similarity}(i,j);$$
$$\text{double upScore} = \text{score}[i][j-1] + \text{similarity}(0,j);$$
$$\text{double leftScore} = \text{score}[i-1][j] + \text{similarity}(i,0);$$

First of all, the first row and column of matrix is filled with zero. Then for rest of cells in matrix, there are three possibilities of score. Suppose a cell x, for which score value can come from three directions (from upper cell, from left cell, or from right cell) and if the value comes from the diagonal cell, then score of cell x will be added to match or mismatch value. If the value comes from upper cell and left cell, then score of cell x is added to gap value.

$$\text{score}[i][j] = \text{Math.max}(\text{diagScore}, \text{Math.max}(\text{upScore}, \text{Math.max}(\text{leftScore}, 0)));$$

Then, find the maximum score value from scoring matrix and check that from which direction the maximum score value of cell x is obtained suing following code.

$$\text{prev\_cells}[i][j]| = \text{dr\_diag};$$

It means the diagonal direction gives maximum score to cell x.

$$\text{prev\_cells}[i][j]| = \text{dr\_left};$$

It means the left direction gives maximum score to cell x.

$$\text{prev\_cells}[i][j]| = \text{dr\_up};$$

It means the upper direction gives maximum score to cell x.

getMaxScore()

This method gets the maximum score from all the scoring matrixes by comparing the all cell values and storing the higher value to MaxScore. After comparing all cell values, the value stored in scoring matrix will be maximum score value in the scoring matrix. Then, this maximum score value which is in integer is normalized and converted to double value.

printAlignments()

This method prints the alignments having maximum score.

If the score comes from left direction, then align the two sequences using gaps as follows:

p = printAlignments($i - 1$, j, qstr1.charAt($i - 1$) + aligned1, "_" + aligned2);

If the score comes from upper direction, then align the two sequences using gaps as follows:

p = printAlignments(i, $j - 1$, "_" + aligned1, str2.charAt($j - 1$) + aligned2);

If the score comes from diagonal direction, then align the two sequences using characters only and without gaps as follows:

$$p = \text{printAlignments}(i-1, j-1, \text{qstr1.charAt}(i-1) + \text{aligned1}, \text{str2.charAt}(j-1) + \text{aligned2});$$

printAlignments(String)

This method extends the alignments obtained with printAlignments() method. These alignments are extended to get higher scores and then again aligned with same procedure before and final alignments we get with the higher scores will be the final suitable alignments. The number of extended alignments we get is the number of hits which we can see in the output at the end as number of matches = 25.

6 Experimental Evaluation

In Fig. 4.5, number of matches varies with varying word size. Threshold value is taken to be 10. With word size 7, number of matches are 16, and when word size = 11 and 15, then number of matches are zero (Figs. 11 and 12).

In Fig. 4.6, number of matches varies with varying word size. Threshold value is taken to be 3. With word size 3, number of matches are 120; when word size is 4, number of matches are 92; and when word size is 5, then number of matches are 74. That means when the word size is increasing, the number of matches has a score greater than the threshold decreases (Fig. 13).

In Fig. 4.6, number of matches varies with varying word size. Threshold value is taken to be 4. With word size 3, number of matches are 0; when word size is 4, number of matches are 58; and when word size is 5, then number of matches are 67.

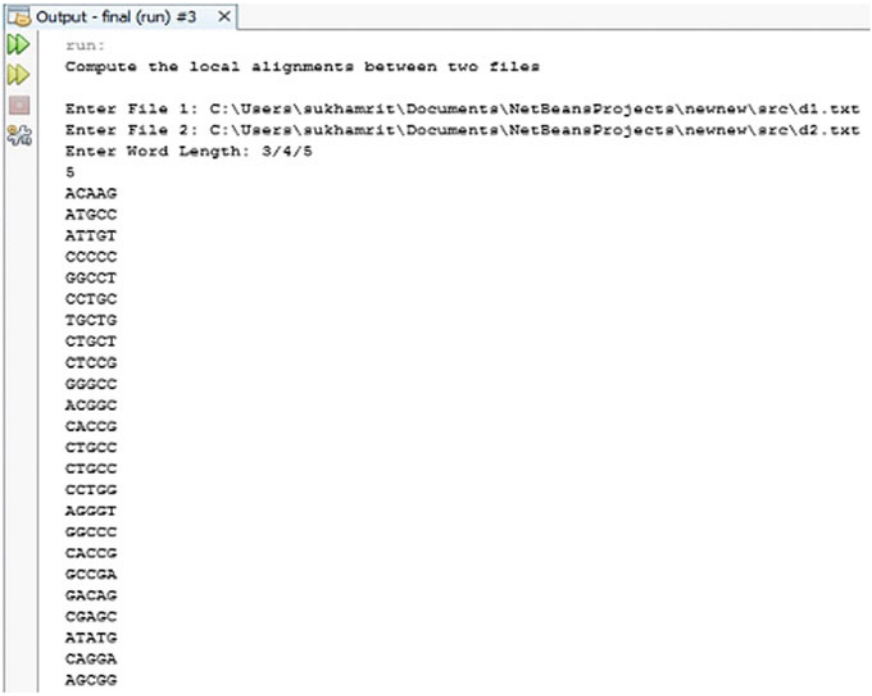

Fig. 11 Segments of query sequence

That means when the word size is increasing, the number of matches has a score greater than the threshold increases because of the threshold (Figs. 14 and 15).

In Fig. 4.6, number of matches varies with varying word size. Threshold value is taken to be 5. With word size 3, number of matches are 0; when word size is 4, number of matches are 0; and when word size is 5, then number of matches are 25. That means when the word size is increasing, the number of matches has a score greater than the threshold increases because of the threshold (Figs. 16 and 17).

7 Conclusion

DNA barcoding is a system for fast and accurate species identification which will make ecological system more accessible. It has many applications in various fields such as preserving natural resources, protecting endangered species. For species identification, sequence alignment is done in somewhat similar sequences.

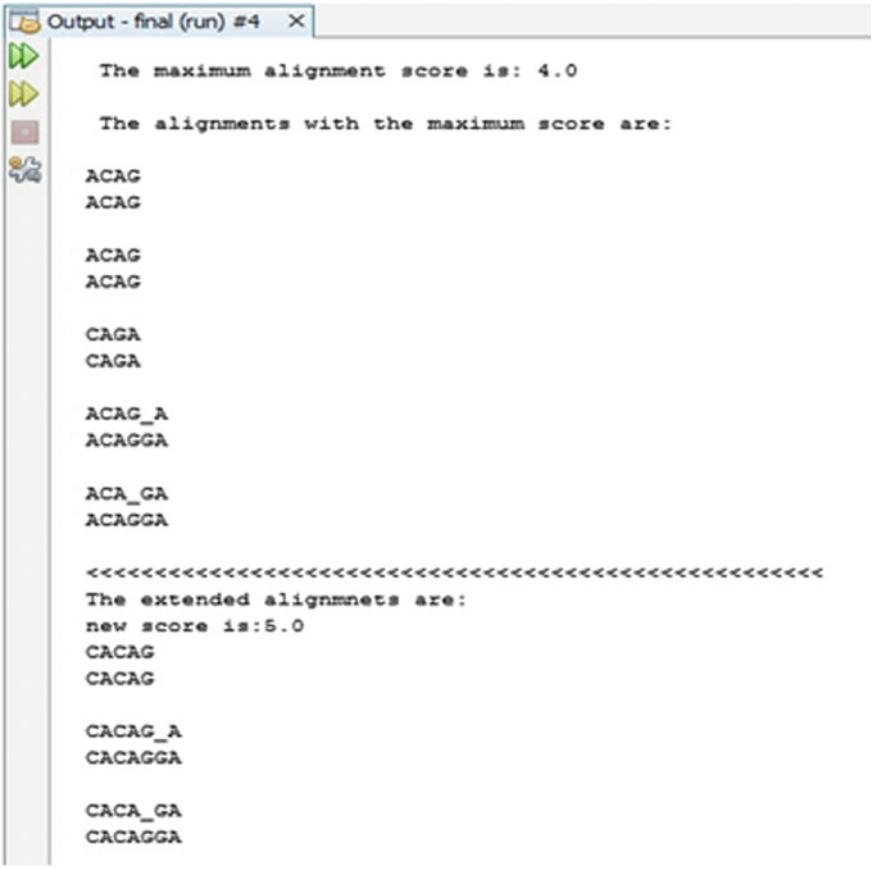

Fig. 12 Extended alignments

Similarity search methods [1] have been used that include some already existing algorithms such as Smith–Waterman, BLAST, and FASTA. BLAST is being used for fast species identification but does not give accurate results like Smith–Waterman which is a very slow process. BLAST has been performed using a sequence to study the effect of word size on accuracy, and the results show that larger the word size, less will be number of hits and vice versa. So the idea is to combine the features of accuracy of Smith–Waterman and speed of BLAST algorithm. An algorithm is proposed and implemented on this idea with word size 3, 4, and 5, having improved accuracy with more number of matches. The number of matches varies with change in word size and the threshold value (Table 3).

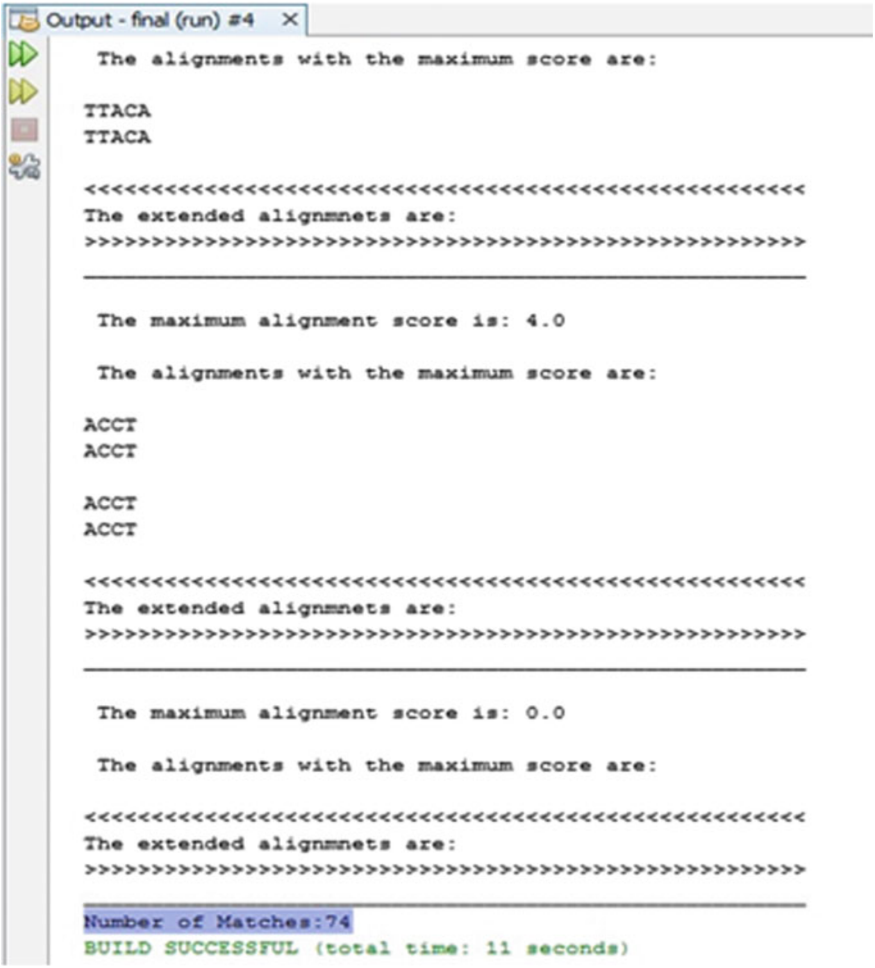

Fig. 13 Number of matches

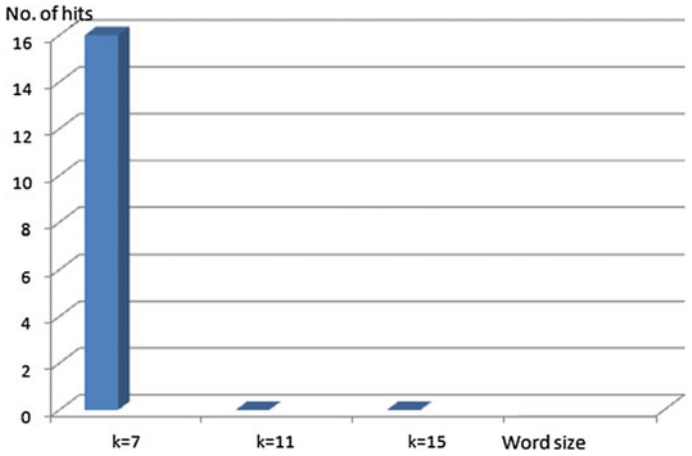

Fig. 14 Number of matches obtained with BLAST with varying word size and threshold = 10

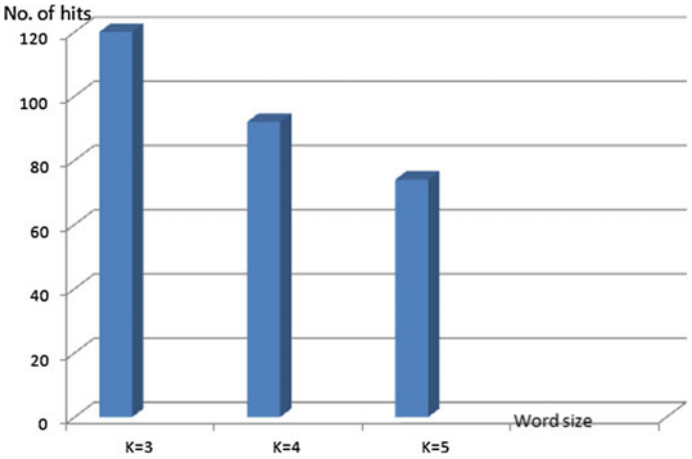

Fig. 15 Number of matches obtained with proposed code with varying word size and threshold = 3

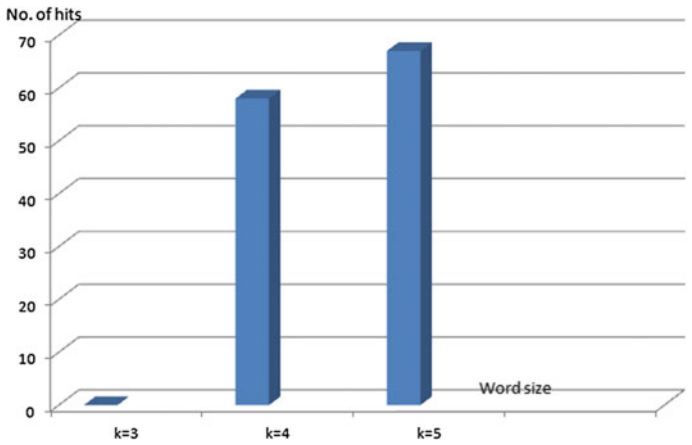

Fig. 16 Number of matches obtained with proposed code with varying word size and threshold = 4

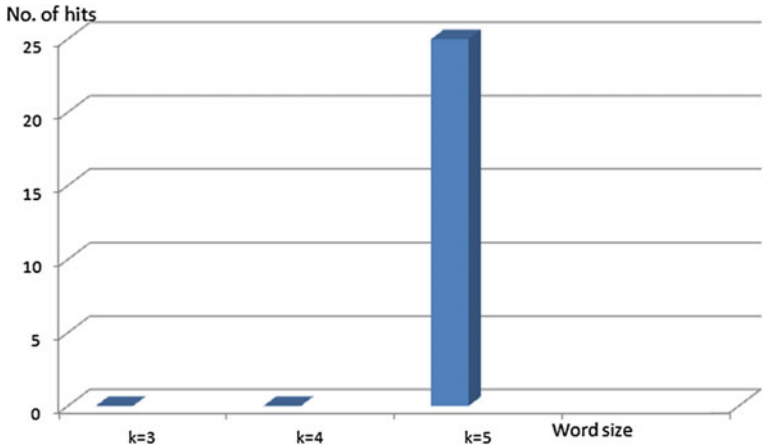

Fig. 17 Number of matches obtained with proposed code with varying word size and threshold = 5

Table 3 Different number of hits with different word size

Word size	Number of hits		
BLAST with threshold T = 10			
7	**16**		
11	**0**		
15	**0**		
Proposed code			
Threshold (*T*) →	3	4	5
3	**120**	**0**	**0**
4	**92**	**58**	**0**
5	**74**	**67**	**25**

References

1. V. Vaidya et al., A review of bioinformatics applications in breast cancer research. J.Adv. Bioinform. Appl. Res. **1**(1), 59–68 (2010)
2. N.M. Luscombe, What is bioinformatics? A proposed definition and overview of the field. J. Methods. Inf. Med. **40**(4), 346–358 (2001)
3. J. Cohen, Bioinformatics—an introduction for computer scientists. J.Comput. Biol. **36**(2), 122–158
4. M.B. Priyadarshni, Applications of bioinformatics. Available at: www.biotecharticles.com (2014)
5. Using DNA barcode to identify and classify living things (2014)
6. K. Sasikumar, C. Anuradha, DNA barcoding as a tool for algal species identification and diversity studies, **7**, 75–76 (2012)
7. P. Hebert, T.R. Gregory, The promise of DNA barcoding for taxonomy, **54**(5), 825–859 (2005)
8. F.C. Pereyra, M.G. Meckan, V. Wei, O. O'Shea, C.J.A. Bradshaw, C.M. Austin, Identification of Rays through DNA Barcoding: An application for ecologists, **7**(6), 1–10 (2012)
9. Identifying species with DNA barcoding. Available: www.barcodeoflife.org (2010)
10. R.D. Collins, R. Hanner, P.D.N. Hebert, The campaign to DNA barcode all fishes, FISH-BOL. J. Fish Biol. **74**, 329–356 (2009)
11. C. Ebach, C. Holdrege, DNA barcoding is no substitute for taxonomy, **434**, 697 (2005)
12. M.C. Ebach, C. Holdrege, More taxonomy, not DNA barcoding, **55**(10), 822–823 (2005)
13. M. Maheshwari, Sequence alignment, in *Introduction to Bioninformatics*, 1st edn. (New Delhi, 2008), pp. 164–196
14. Available at: www.dnabarcodes.org
15. W.J. Kress, D.L. Erickson, Introduction, in *DNA Barcodes Methods and Protocols* (Washington, DC, USA, 2012), pp. 3–10
16. P.M. Hollingsworth, Anti intellectualism in DNA barcoding enterprise. Zoologia. **3**(2), 44–47 (2007)
17. K.K. Dasmahapatra, J. Mallet, DNA barcoding: recent successes and future prospects. Hered. **97**, 254–255 (2006)
18. R.A. Collins, R.H. Cruik Shank, The seven deadly sins of DNA barcoding. Mol. Ecol. Res. 1–7 (2012)
19. C. Moritz, C. Cicero, DNA barcoding: promise and pitfalls. PLos Biol. **2**(10), 1529–1531 (2004)
20. P. Hebert, A. Cywinska, S.L. Ball, J.R. deWaard, Biological identifications through DNA barcodes. Proc. Biol. Sci. **270**(1512), 313–321 (2003)
21. C. Notredame, Recent evolutions of multiple sequence alignment algorithms. PLoS. Comput. Biol. **3**(8), 1405–1408 (2007)
22. D.P. Little et al., A comparison of algorithms for the identification of specimens using DNA barcodes: examples from gymnosperms. PLoS ONE. 1–21 (2007)
23. J. Waugh, DNA barcoding in animal species: progress, potential and pitfalls. Bio Essays. **29**(2), 188–197 (2007)
24. D.L. Dalton et al., DNA barcoding as a tool for species identification in three forensic wildlife cases in South Africa. Forensic. Sci. Int. **207**(1–3), 51–54 (2011)
25. Z.T. Nagy, T. Backeljau, M.D. Meyer, K. Jordaens, (2013), DNA barcoding: a practical tool for fundamental and applied biodiversity research. Zookeys. **5**(24)
26. C.B. Abilash, K. Rohitaktha, A comparative study on global and local alignment algorithm methods. IJETAE. **4**(1), 34–43 (2014)
27. A.G. Khallaf et al., DNA barcoding reveals a high level of mislabelling in Egyptian fish fillets. Food control. **46**, 441–445 (2014)

28. M.S. Rosenberg, Sequence alignment, in *Sequence Alignment Methods, Models, Concepts And Strategies*, 1st edn. (California, 2009)
29. D.E. Krane, M.L. Raymer, Data searches and pairwise alignments, in *Fundamental Concepts of Bioinformatics*, 1st edn. (New Delhi, 2006)
30. B. Bergeron, Pattern matching, in *Bioinformatics Computing* (New Delhi, 2003), pp. 302–339
31. Nucleotide Database, NCBI, NLM, NIH, 8600 Rockville Pike, Bethesda, MD 20894
32. EMBL-EBI, Pair wise sequence alignment, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire, CB10 1SD, UK
33. By blast-help group, NCBI User Service, BLAST program selection guide. NCBI, NLM, NIH, 8600 Rockville Pike, Bethesda, MD 20894
34. NCBI News, GenBank surpasses one trillion total bases of publicly available sequence of data. NCBI, NLM, NIH, 8600 Rockville Pike, Bethesda, MD 20894 (2015)
35. J. Hanken, Trends Ecol. Evol. **18**(2) (2003)
36. Available at: http://seqcore.brcf.med.umich.edu/
37. H.A.I. Ramadan, N.A. Baeshen, Biological identification through DNA barcodes (2012)