International Journal for Research in Technological Studies/ Vol. 5, Issue 2, January 2018 / ISSN (online): 2348-1439

Identification of Organisms using DNA barcodes

Sandeep Kaur¹ Amandeep Kaur² ^{1,2}Assistant Professor

^{1,2}Department of Computer Science & Engineering

¹GNDU RC Fattu Dhinga, Sultanpur Lodhi, Punjab, India ²GNDU RC Gurdaspur, Punjab, India

Abstract— DNA bar-coding is a system for fast and accurate species identification which will make ecological system more accessible by using short DNA sequence instead of whole genome and used for eukaryotes. The short DNA sequence is generated from standard region of genome known as marker. DNA bar-coding has many applications in various fields like preserving natural resources, protecting endangered species, controlling agriculture pests, identifying disease vectors, monitoring water quality, authentication of natural health products and identification of medicinal plants. For species identification using DNA bar-coding, similarity search methods have been used that include some already existing algorithms like Needleman Wunsch, Smith waterman, BLAST and FASTA. BLAST is being used for fast species identification but not give accurate results like Smith waterman but it is a very slow process. BLAST has been performed using a sequence to study the effect of word size on accuracy and results show that larger the word size, less will be number of hits and smaller the word size, more will be number of hits. More number of hits means more accuracy. So idea is to combine the features of accuracy of Smith Waterman and speed of BLAST algorithm and a algorithm is proposed with combined features of both algorithms. Keywords— DNA, marker, barcode

I. INTRODUCTION

Monitoring the biological effects of global climate, Identification of organisms has become important to preserve species because of increasing habitat destruction. There is estimation of 5 to 50 million plants and animals, living on earth, out of which less than 2 million have been identified. Extinction of animals and plants is increasing yearly means thousand of them are lost each year and most of them are not

identified yet.[1] The high level of destruction and endangerment of ecosystem has lead to improved system for identifying species. In recent years new ecological approach called DNA bar-coding has been proposed to identify species and ecology research. [2][3] DNA bar-coding is a system for fast and accurate species identification which will make ecological system more accessible. [4] It 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 bar codes". DNA bar-coding is a new tool for identification of species and for taxonomic research. It is not a new concept as Carl Woese used rRNA and molecular markers like rDNA and mtDNA to discover archea i.e. prokaryotes and then for drawing evolutionary tree. But DNA bar-coding uses short DNA sequence instead of whole genome and used for eukaryotes. This short DNA sequence is taken 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 non experts can identify species from small, damaged or industrially processed material. [5]



Fig.1. DNA Barcode[15]

The standard region used to generate DNA barcode is known as marker. This marker varies among species. In animals Paul Hebert proposed the use of COI 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. Reason for choosing this is because of its small intraspecific and large interspecific differences. It is not suitable for other group of organisms because it is uniform in them. So ITS (Internal Transcribed Spacer) is recognized for fungus and two genes from chloroplast genome, rbcl and matK are recognized as barcode markers for plants by IBOL. [12][13]

The sequence data generated from standardized region is used for identification of organism and to construct a phylogenic tree. In this tree related individuals are clustered together and can provide large amount of information about specie. [11][14].

1.1 Applications of DNA Bar-coding

A. Controlling Agricultural Pest

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

B. Identifying Disease Vectors

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

C. Sustaining Natural Resources

Over harvesting of natural resources like hardwood trees and fishes is causing species, extinction and economies collapse of industries that rely on them. Using DNA barcoding natural resource managers can monitor illegal trade of products made of these natural resources. Fishbol is reference barcode library for hardwood trees to improve management and conservation of natural resources.

D. . Protecting Endangered Species

Primate Population is reduced by 90% in Africa because of bush meat hunting. DNA bar-coding can be used by law enforcement to bush meat in local markets which is obtained from bush meat.

E. Monitoring Vector 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 bar-coding is used to create a library of these species that can be difficult to identify. Bar-coding can be used by environmental agencies to improve determination of quality and to create better policies which can ensure safe supply of drinking water.

F. Routine Authentication of Natural Health Products

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

G. Identifying of plant leaves even if flowers or fruit are not available

H. Identification of medical plants

1.2 Procedure of DNA bar-coding

The process of DNA bar-coding involves two basic steps: First is building the barcode library of identified species and second is matching the barcode sequence of the unknown sample with the barcode library (known as sequence alignment) for its identification. The first step requires ecologic expertise in selecting one or several individuals per species as reference samples in the barcode library. Tissue samples for generation barcodes are either housed in museum or they can be live specimen in the field. These specimens go through lab processes that are tissue sampling and DNA processing and sequencing to generate DNA barcode in form of chromatogram. Chromatogram is visual representation of DNA sequence produced by sequencer. This barcode can be stored in database for future use or can be used as query sequence to be compared with sequence already present in database. [6]



Fig.2. DNA Barcoding Procedure



II. 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. [8][11] 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 means these need to be compared with sequences in database. 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 identifying the new species, finding relationship between species and for predicting the function and structure of genes and proteins. [7]

- DNA sequence alignment is of two types:
 - a) Pair wise sequence alignmentb) Multiple sequence alignment

Pair wise sequence alignment is a process of aligning or comparing two sequences at one time. Multiple sequence alignment is process of aligning or comparing more than two or three sequences with database sequences for doing phylogenetic analysis. It is done to study and analyze the relationship between various taxa using phylogenetic or evolutionary tree. We will consider pair wise alignment in our paper.

III. PAIR WISE SEQUENCE ALIGNMENT

Pair wise 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 aligned, and if these two sequences are 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. [9] 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. [10] 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. Local alignments are appropriate for dissimilar sequences which may contain similar character sequence. [9]

A. Already Existing Pair Wise Sequence Alignmnet

Some already existing algorithms for pair wise sequence alignment are Needleman-Wunsch, Smith Waterman, FASTA (Fast Alignment) and BLAST (Basic Local Alignment Search Tool).

1) Needleman-Wunsch Algorithm

It 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 following.

A and B are sequences and Ai and Bj represents the base 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, you 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.

The time complexity of this algorithm is O(MN) and space complexity is also same i.e. O(MN).

2) Smith Waterman Algorithm

It was published in 1981. The Smith–Waterman algorithm is a well-known algorithm used for local sequence alignment. It is very similar to Needleman-Wunsch algorithm only difference is that instead of looking at the total sequence, the Smith–Waterman algorithm compares segments of all possible lengths and checks for similarity. Algorithm is as following:

Stop

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 encounter the value "0" in the score matrix. [9]

The time and space complexity is same as of Needleman Wunsch algorithm. Space complexity is same because same matrix is used and same amount of space for trace back is needed.

3) FASTA

FASTA stands for fast alignment. FASTA is fast searching algorithm used for comparing query sequence with database. It comes under dynamic programming 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. Algorithm is as following:

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 size of the k-tuples, that means larger the k-tuples, the faster the algorithm. The true complexity is not easily determined because the speed of alignment of two sequences depends on total number 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 use less space because not all cells in the matrix are marked.

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 regions of high local similarity in alignments without gaps. Algorithm:

Step 1: Word Search Method: Sequence is filtered to remove complexity regions

Step 2: Identification of exact word match method, searches the database for neighbourhood word. Words having equal or greater scores than neighbourhood score threshold are taken for alignment.

Step 3: 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 bar-coding has been used as a tool for identification of three species in forensic wildlife in South Africa [16] and also it has revealed high level of mislabelling in fish fillets purchased from Egyptian markets. [17]

IV. METHODOLOGY & COMPARISION

Parameters to be considered are word length or word size and number of hits in 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 sequences.

To study the effect of change in work length on accuracy of species identification, BlastN was performed using mRNA sequence (nucleotide sequence) of invertebrate animal species named as Anaspides Tasmania against non redundant database. It returns top 100 sequences having some similarity, for each query sequence.[18] It is a fresh water 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 were used with expect value, E=10.

Sequence used for the observation is as follows which is extracted from CO1 region of anaspides specie and is of size 657 bases and is in Fasta format.

>EMBOSS_001

TCTTTAGATTTTATTTTTGGAGCTTGGTCTGGCATA GTAGGCACCGCCCTAAGACTTATTATTCGGGCTGA ATTAGGACAACCTGGTAGACTTATTGGTGATGATC AAATTTACAACGTGGTCGTAACAGCTCATGCTTTTG TGATAATTTTTTTTATAGTTATGCCCATTATAATTG GTGGATTTGGAAATTGACTTGTTCCCTTAATATTAG GTGCTCCTGATATAGCTTTTCCTCGTATAAATAATA TAAGATTTTGACTTCTTCCACCTTCTTTAACTCTTCT CCTATCCAGAGGAATAGTTGAAAGAGGTGTTGGCA CAGGATGAACTGTTTATCCTCCTTTAGCTGCTGGAA TCGCCCATGCAGGCGCTTCTGTGGACTTAGGAATTT TTTCTCTTCATATAGCGGGAGCTTCTTCTATTAG GGGCGGTAAATTTTATTACTACTTCTATTAATATGC GTGCCAATGGTATAACTTTAGATCGAATACCTTTAT TTGTCTGATCCGTTTTATTACTGCTATTCTTTACT ACTCTCTCTCCCGTTTTAGCAGGGGCAATCACAAT ACTTCTCACTGACCGTAACTTAAATACTTCTTTCTT TGACCCCGCTGGAGGAGGAGATCCATTCTTTATCA ACATAAATGCC

 Table 3.1 Varying number of hits with different word size

	Word	No. of hits
4	size	
	(k)	
	7	518310295
1	11	32757086
	15	14769504

The results from word size k=7 returned 518310295 hits, k=11 returned 32757086 i.e. less hits than returned by k=7 and then k=15 returned 14769504 which is least of all. So the observations tells 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. [19]



Fig.4. Comparison on the basis of word size Table 3.2 Comparison of Sequence alignment algorithms

	Complexi ty	Alignme nt	Accurac y	Speed
Needlem an Wunsch	O(MN)	Global alignmen t	Less accurate than smith waterma n	Slow for searchin g entire databas e
Smith waterman	O(MN)	Local alignmen t	More accurate	Slow for searchin g entire databas e

FASTA	Time complexit y depends on k	Local alignmen t	Less accurate than Smith Waterm an	Faster than above
BLAST	O(MN)	Local alignmen t	Less than Smith Waterm an	Fastest

V. PROPOSED WORK

In this paper, an algorithm is proposed for local sequence alignment which gives more accurate results for better sequence alignment.



Fig.5. Design of Proposed Work

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

ACTGACTGCCCGTAAATGCATC

Then from the indexed table, the words are matched with sequences present in database. The databases used for DNA barcode are BOLD and Genbank. 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. Then the highest scored pair is chosen.

A. 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. These scores are calculated using a substitution matrix by including gaps in sequences. These gaps are also known as indels (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 words.

Step 5: Extend the hit (matched word) in both directions until its score is increasing.

Step 6: Report the highest scoring pair if its score is greater than cut off and lower than expect value.

B. Feature in Proposed Algorithm.

Proposed algorithm is combination of best features of smith waterman and BLAST algorithm. That means accuracy of identification provided by smith waterman and fast search provided by heuristic technique of BLAST algorithm. So, proposed algorithm provides faster search and accurate results.

Also word size used will be 3 to 5 for faster search and sensitivity (accuracy). Because speed is directly proportional to word size and sensitivity is inversely proportional to word size. So, 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 to 5.

C. Parameters

1) Word Size

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

2) Threshold

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

3) Expect value

It is the number of hits one can expect that means estimation of how many times you would expect a result. Its default value to be used is 10.

4) Cut off value

It is used for reporting Highest Scoring Segment. Its default value to be used is calculated from expect value.

VI. CONCLUSIONS

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 like preserving natural resources, protecting endangered species. For species identification similarity search methods [20] have been used that include some already existing algorithms like Neeldeman Wunsch, Smith waterman, BLAST and FASTA. BLAST is being used for fast species identification but not give accurate results like Smith waterman but it is a very slow process. BLAST has been performed using a sequence to study the effect of word size on accuracy and results show that larger the word size, less will be number of hits and vice versa. So idea is to combine the features of accuracy of Smith Waterman and speed of BLAST algorithm. An algorithm is proposed on this idea with word size 3 to 5, which will have more accuracy and more speed.

REFERENCES

- [1] Using DNA barcode to identify and classify living things. 2014
- [2] Z.T. Nagy, T. Backeljau, M.D. Meyer and K. Jordaens (2013), DNA barcoding: A practical tool for fundamental and applied biodiversity research, 5(24)
- [3] John Waugh (2007), DNA barcoding in animal species: progress, potential and pitfalls, 29, pp 188-197.
- [4] Paul Hebert and T.R. Gregory (2005), The promise of DNA Barcoding for taxonomy, 54(5), pp 825-859.
- [5] (2010). Identifying species with DNA barcoding. Available: www.barcodeoflife.org
- [6] W.J. Kress, D.L. Erickson, "Introduction", in DNA barcodes methods and protocols, Washington, DC, USA, 2012, pp 3-10.
- [7] D.E. Krane and M.L. Raymer, "Data searches and pairwise alignments", in Fundamental concepts of bioinformatics, 1st ed., New Delhi, 2006.
- [8] M.S. Rosenberg, "Sequence alignment", in Sequence alignment methods, models, concepts and strategies, 1st ed., California, 2009.
- [9] Abilash CB. And Rohitaktha K (2014). A comparative study on global and local alignment algorithm methods, 4(1), pp 34-43.
- [10] Bryan Bergeron, "Pattern matching", in Bioinformatics computing, New Delhi, 2003, pp 302-339.
- [11] Meetu Maheshwari, "Sequence alignment", in Introduction to bioninformatics, 1st ed., New Delhi, 2008, pp 164-196.
- [12] K. Sasikumar and C. Anuradha (2012). Dna barcoding as a tool for algal species identification and diversity studies, 7, pp 75-76.
- [13] C. Ebach and C. Holdrege (2005), Dna barcoding is no substiture for taxonomy, 434, pp 697.
- [14] M.C.Ebach and C.Holdrege (2005). More taxonomy, not DNA barcoding, 55(10), pp 822-823.
- [15] J. Hanken. (2003, Feb.). Trends in ecology and evolution. 18(2).
- [16] D.L.Dalton et al. (2011), DNA Barcoding as a tool for species identification in three forensic wildlife cases in south Africa, pp 51-54.
- [17] A.G. Khallaf at al. (2014), DNA Barcoding reveals a high level of mislabelling in Egyptian fish fillets, pp 441-445.
- [18] D.P.Little et al. (2007), A comparison of algorithms for the identification of specimens using DNA Barcodes: examples from gymnosperms, pp 1-21.
- [19] By blast-help group, NCBI User Service, "BLAST Program Selection Guide", NCBI, NLM, NIH, 8600 Rockville Pike, Bethesda, MD 20894.
- [20] Available at: http://seqcore.brcf.med.umich.edu/