International Journal of Electrical and Computer Engineering (IJECE)
Vol. 6, No. 3, June 2016, pp. 1133 ~ 1139
ISSN: 2088-8708, DOI: 10.11591/ijece.v6i3.10086

DNA Bar-Coding: A Novel Approach for Identifying an Individual Using Extended Levenshtein Distance Algorithm and STR Analysis

Likhitha C. P, Ninitha P, Kanchana V
Department of Computer Science, Amrita Vishwa Vidyapeetham University, Mysuru campus, Karnataka, India

Article Info

Article history:

Received Feb 24, 2016
Revised May 11, 2016
Accepted May 28, 2016

Keyword:

Color DNA bar-code
DNA bar-coding
Human identification
Levenshtein distance algorithm
Sequence matching
STR

ABSTRACT

DNA bar-coding is a technique that uses the short DNA nucleotide sequences from the standard genome of the species in order to find and group the species to which it belongs to. The species are identified by their DNA nucleotide sequences in the same way the items are recognized and billed in the supermarket using barcode scanner to scan the Universal Product Code of the items. Two items may look same to the untrained eye, but in both cases the barcodes are distinct. It was possible to create DNA-barcodes to characterize species by analysing DNA samples from fish, birds, mammals, plants, and invertebrates using Smith-waterman and Needleman-Wunsch algorithm. In this work we are creating human DNA barcode and implementing Extended Levenshtein distance algorithm along with STR analysis that uses less computation time compared to the previously used algorithms to measure the differential distance between the two DNA nucleotide sequences through which an individual can be identified.

Corresponding Author:

Likhitha C P,
Department of Computer Science,
Amrita Vishwa Vidyapeetham University, Mysuru campus,
#114, 7th Cross, Bogadi 2nd Stage, Mysuru-570026.
Email: likhiponnappa7@gmail.com

1. INTRODUCTION

DNA bar-coding is a framework for quick and accurate species recognition that makes ecological system more available by using short DNA sequence rather than entire genome. The short DNA sequence is produced from standard region of genome known as marker. This marker is different for different species like CO1 cytochrome c oxidase 1 for creatures, matK for plants, internal transcribed spacer (ITS) for fungus and mitochondrial gene for humans. DNA bar-coding has numerous applications in different fields like preserving natural resources, securing endangered species, recognizing disease vectors, identifying agricultural pests, identification of medicinal plants and identification of humans.

As of recently, biological species were recognized using morphological elements like the shape, size and shade of body parts. In many situations, an expert could make routine distinguishing pieces of proof using morphological "keys" (step-by-step instructions of what to search for, however much of the time an accomplished proficient taxonomist is required. On the off chance that if a species is damaged or is in an immature phase of improvement, even a proficient taxonomist may be not able to identify and distinguish that species [1]. Bar coding take care of these issues in light of the fact that even non-authorities can get standardized identifications from small measures of tissue. This is not to say that traditional scientific classification has turned out to be less important. Maybe, DNA bar coding can fill a double need as another device in the taxonomist's tool stash supplementing their insight and in addition being a creative gadget for non-specialists who need to make a quick recognition.

Until now, DNA bar-coding technique was proved useful in identifying species of insects [2], fishes [3], Canadian mosquito [4], spiders [5], birds [6] and animals [7]. It was also used effectively to examine Hyalella [8], a taxonomically difficult genus of amphipod crustaceans and tussock moths (Lepidoptera: Lymantriidae) [9].

Human DNA bar-coding is a powerful tool in forensics to identify the human [10] through the DNA samples stored in the database. This works by collecting the sample DNA sequence from the individual, converting this sequence into color barcodes based on the nucleotide bases [11] and storing these barcodes in the standard library along with the complete details pertaining to that particular individual. By scanning the barcode [12] or by entering the sequence, the newly entered sequence is compared to the stored sequence in the library and matched. If the sequence match is found, the complete details of the matched sequence are displayed to the user. This tool has many applications in the areas like identifying criminals whose DNA may match evidence left at crime scene, to exonerate persons wrongly accused of crimes and to establish family relationships. Human DNA bar-coding include different activities such as

a) Working with the individuals: To collect, identify, classify and store individuals' data in secure repositories.
b) Barcode generation: Color DNA barcode of the individual sequences is generated by pre-processing the nucleotide sequences.
c) Managing data: The generated color DNA barcodes along with the sequences and the details regarding the individual are updated in the standard library.
d) Finding the match: Either DNA sequences or barcode images are uploaded to match with the database to display the details of the matched sequence.

Previously Smith-waterman [13] and Needleman-Wunsch algorithm [14] were used in sequence alignment. To increase the efficiency, the Levenshtein distance algorithm [15] is implemented to compute the number of mismatches between two long strings.

In few cases, the Levenshtein distance algorithm may return same mismatch count for multiple sequences. Hence to find the exact match between those sequences the Short Tandem Repeats analysis (STR) [16],[17] is used. STR analysis is a tool used in forensic to analysis and evaluate the specific STR regions [18] that is on nuclear DNA. The STR regions that are analyzed from nuclear DNA may have polymorphic nature. But in the forensic testing of these STR regions, it shows differentiation between one DNA profile and another.

2. RESEARCH METHOD

2.1. Bar-coding

As each individuals fingerprint is different, each individuals DNA is also different. By DNA bar-coding we can identify species or individuals fast and accurately. Same as fingerprinting technology, DNA bar-coding can also help in finding out the culprit in the criminal cases or unidentified victim. The biological samples are collected from Blood, Saliva, Urine, Hair, Bone or Tissue of an individual and sent to laboratory to extract DNA sequence. The sequence contain four nucleotide bases A-adenine, T-thymine, G-guanine and C-cytosine. These four nucleotide random combination leads to large sequences of DNA with varying lengths. For example,

1. (ATTCAAAAGACCTCGCTAAAAATCTCGCAGTCAACTATCTTTAGCGTTAAATCACGCAACA TATTTCAACCGCATTGGAGAGTCGAGGCAGCTAAGCCCGGTAACCCCTTTCATATCTGATCC TACGGGATCTTGGGTTTGTCCGCCATTCTGATTGTGAGAACGGGGTGTGTCCGCAGAACCCC TCTCTAGACAACTAGACCATTCGACTCAG),
2. (TCGAGAATAAAAGTTTCAGTGTAATAAACCAAGATGTCTTATCTGACGCGAGCTTCCTTCT TTGAAGTAACAGTTCCTGTCTCGTCTTCACTAAATCTTCACAGCGCGTCCTAATACCGGCAG TGAACCGTATCCGGTTACTATATGCTGTTGTTAGAGCGTTCTCGCACGCGACATTACAGTAC CTCGCCCAGTCGCAATTCTGCCTGC) and so on.

The DNA sequence is bar-coded and saved along with all details of the individual in the database. The DNA is bar-coded by assigning the four nucleotides with different colors as follows: A - Green, T - Red, G - Black, and C -Blue.

The following is a part of DNA sequence taken as sample and bar-coded as shown in Figure 1.
GTTGAAGCGGTTATCGCGCAAAAAAGCTGGCGCCCGGAGAGTGGCATGCAAAGCTGTCAGCAA ACCCAACGTTGATCAACGCAGCGCAGCTTGAGTGTCTTTCTTTGGCCCATACCCAGCCCGTGCA ATGACCAACGCGTTAGATTGACCTAGT.

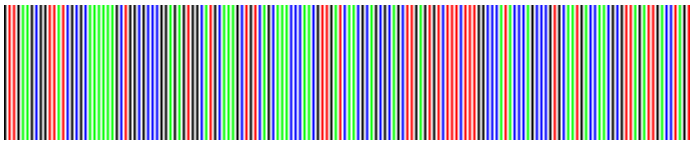

Figure 1. Color DNA bar-code

The database contains millions of records of the individuals with their personal details, DNA sequence and bar-coded image of DNA sequence. If any unidentified victim is found, DNA sequence can be extracted from that individual and given as input to find the match and to retrieve the details of the victim from the database using Extended Levenshtein distance algorithm with STR method.

Information such as name, identification number (ID), reason of death etc. are included in the barcode for the reference. The barcode library provides a function to encode the content into an image which can be saved in JPEG, GIF, PNG or Bitmap formats, and also a function to decode an image. When the individual is identified, the generated color barcode image can be easily printed and attached to the victims sample for further processing instead of carrying out the laboratory process again. In future, if any details about the victim are required, the attached bar-code image of that victim is scanned and inputted to the system.

2.2. Extended Levenshtein distance algorithm

The Extended Levenshtein distance algorithm computes the distance between the two strings. In other words, it computes the number of mismatches between two strings using dynamic programming approach. The two strings used are nucleotide sequences with varying length of bp (base pairs).

The algorithm compares the two sequences and returns number of mismatch between them as shown in Figure 2. It executes until it compare the inputted sequences by the user with all the list of nucleotide sequences stored in the database. Once it compares all the sequences in the database, it returns the minimum mismatch count of the two sequences and its corresponding ID. If the mismatch count is 0 for any DNA sequence, then it is considered exact match and we can retrieve the information using the corresponding ID.

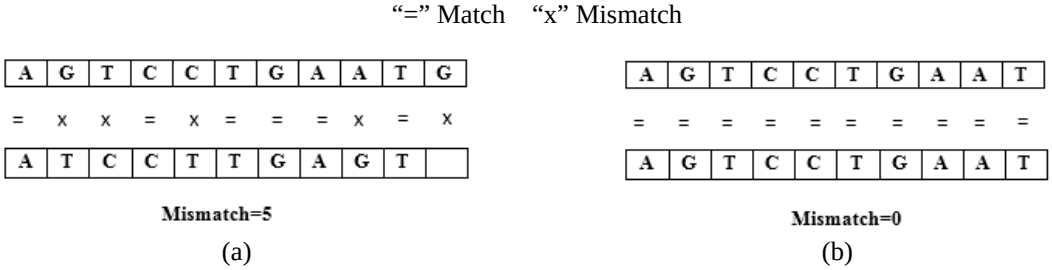

Figure 2. Extended Levenshtein distance algorithm for matching two DNA sequence (a) mismatch count for different string length (b) mismatch count for same string length

2.3. Short tandem repeats (STR)

In few cases, a part of DNA sequence of a person is similar to that of another person. In such situations, it is difficult to match the sequence and predict the individuals. By implementing Extended Levenshtein distance algorithm, the lowest dissimilarity between sequences is obtained. The sequence having lowest dissimilarity other than 0 is analyzed using STR method. STRs are nothing but short sequence of DNA of length 4-5 base pairs that are repeated many number of times in a single nucleotide sequence. STR is used to compare specific loci on nucleotide sequence from two or more samples. STR analysis measures the exact number of repeating units of nucleotide in DNA sequence.

For example, if group of four nucleotides (tetramer) is considered as repeat units, then the number of repeat units in a single genomic sequence is calculated and compared with the number of same group of four nucleotide repeat units in another sequence. If both the count of repeat units of DNA sequences matches then we come to the conclusion that the sequence is of same person.

Extended Levenshtein Algorithm with STR method

Step 1: Read S1
Step 2: for S2 (k to t from database)
Step 3: Check each character of S1 (i from 1 to n).
Step 4: Check each character of S2 (j from 1 to m).
Step 5: If S1[i] equal to S2[j], the mismatch is 0.
If S1[i] not equal to S2[j], the mismatch is 1.
Step 6: k=k+1 goto Step 2
Step 7: Minimum mismatch and sequence ID is obtained.
Step 8: If minimum mismatch=0 then goto Step 13 else goto Step 9
Step 9: Read repeat units.
Step 10: Check the number of repeat units in S2[j].
Step 11: Check the number of repeat units in S1[i].
Step 12: If S1[i] equal to S2[j], match found.
If S1[i] not equal to S2[j], match not found.
Step 13: Print the details whose match found through sequence ID.

3. RESULTS AND ANALYSIS

When a new sequence was entered, it was made to check with the sequences already stored in the database. Then it displayed the number of mismatches along with the ID of the individual for each of the sequence compared as shown in Figure 3.

The end result showed the minimum mismatched count through which the individual was identified using the sequence ID. If the minimum mismatch count = 0, we can conclude that the both sequence have exact match and it belongs to that individual of that ID.

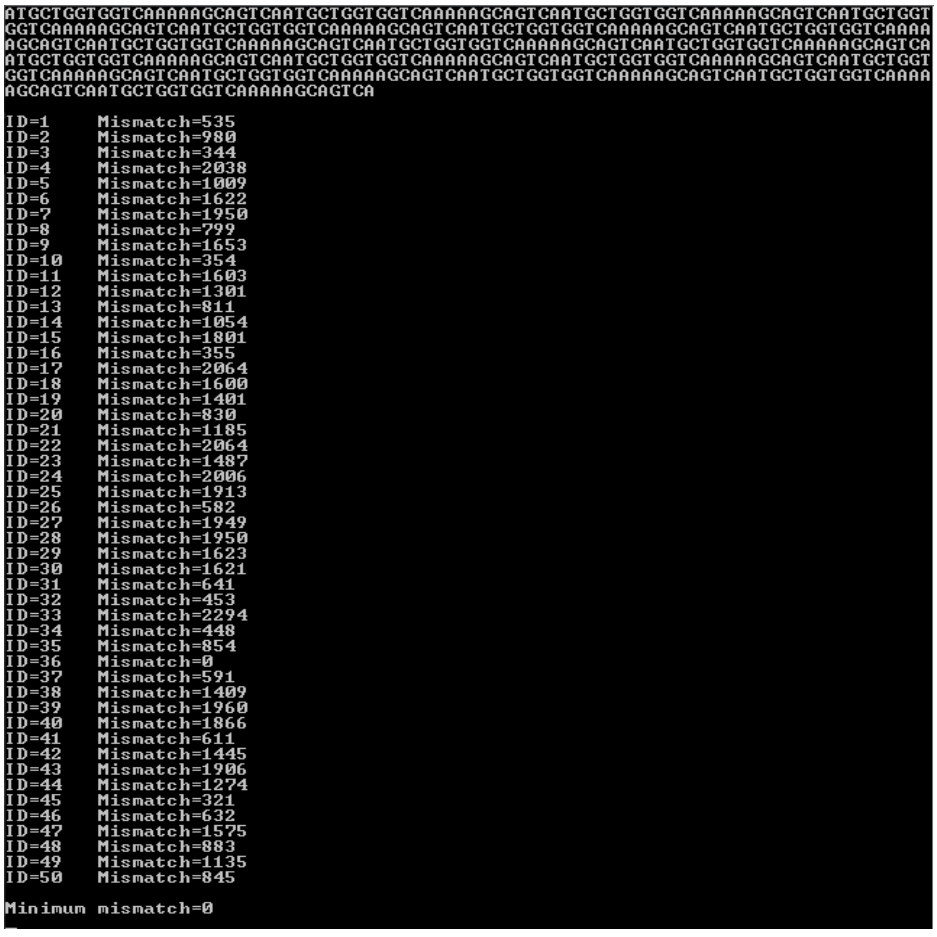

Figure 3. Result of sequence matching

The Figure 4 shows how the individual is identified. The x-axis represents DNA sequence ID of the individual. The y-axis represents minimum mismatch count of the sequence. The graph is plotted for the values obtained in Figure 3, when the sequence is compared using Extended Levenshtein distance algorithm. The point which touches minimum mismatch count 0 is identified as exact match and that sequence ID details is obtained.

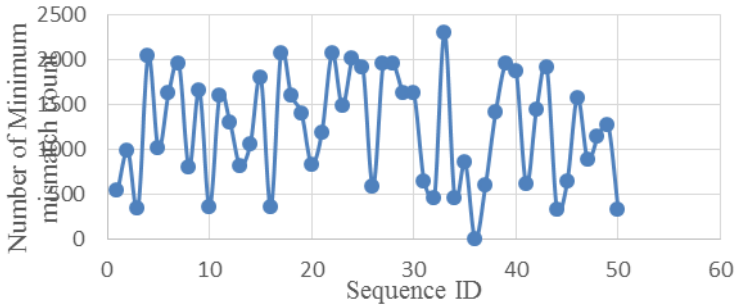

Figure 4. Graphical representation of sequence matching

If the minimum mismatch count is greater than 0 or if in case many IDs exhibit the same minimum mismatch counts, the STR method is used to count and compare the tetramer repeat units in the sequence as shown in Figure 5.

CCTG Repeat units

ACAGTCGCAACACCCCTGACCTGTATTTAGTTAGTTCCGTGTGCCCATGGTGCCTGTATTGTC
TCTATATAAGGACAGATAGGGTTACGTTGCCAATCCCCCCCTGGGCGACAGCGAACGACGCT
AATCGACAGATCCATACTCAGGCCGTACACCCTGTCTTACAAGGAATGATTAGCTAGAGGTC
CGCAACCAAAGACGTGGTAGGGTTCTGCGACAGGTTATGTCCAATGAGTTTCTTGCAGGAGC
CGTGATTATCCTGAGAAGGCCGCTACCTTAGTATGAAGCGATCGTAAACACCCGATACTGGT
TAGGCATCAAAGCACACTCACCCTGATGGCAAATTATGATGTCCAAACTAGGGGCCGAGAG
GTGGACGATACGGTATACGTCATAGCCCACTCGAATTGACATACGCTTGTAGCGCAGTGCTC
TAAGTAAGGATGAATCCTCGGTCGGACACCTGTCCTGCTAAGTTTTCATGCGAGAAACTATA
GGATCAAACGGTTAGTCAACGAGCTCGGATTAGGAGACTTGTTGGACTTCGCAGGATGCAG
CTTTCACTTAATGACACAAGGACGAAGAGATAAACCTTCATAGCTGACTAATAGGTTGGCGG
ATTGAAACCACTTCATTCCTGACATAGTGAAGGTCACCAATCCACTGTTATACCGATTGCAC
TTGACTAATCTTTGCCAGAAACGTGAAGCGCGCTCGGCTAGACGCTTAACGCGGAAAGCACC
GGTTGTTAACGCACTGCAAAGCTCCGCTGTAGCAACGCTTTACAGCTCGAACGTATGAAGGC
TTTCAATATCGGTCCCCAACATTTCCTCTCACATCTGACCAGCTGATCATCGGCCTGAAAGCA
ACGACCCAA

Figure 5. STR analysis

If the count of tetramer repeat units of one sequence is equal to another sequence, we can conclude that both the sequence belongs to the same individual. The entire details pertaining to that person is display with corresponding ID stored in the database.

The analysis proved that the proposed algorithm is efficient as it took less execution time to compare the biological sequences when compared with the smith-waterman and Needleman-wunsch algorithm as shown in Figure 6.

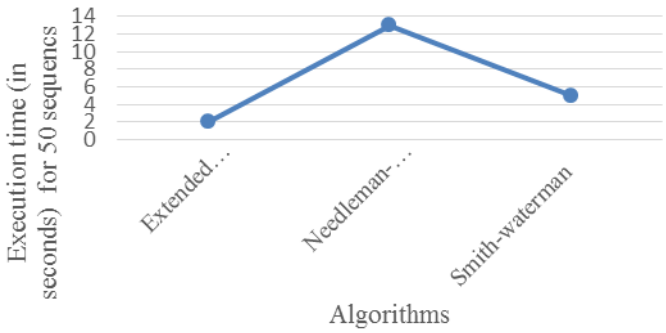

Figure 6. Comparison of algorithms

4. SUMMARY AND CONCLUSION

The *Extended Levenshtein distance algorithm* along with STR analysis is implemented to identify people by finding the least number of mismatches between the sequences through DNA bar-coding. This helps the government organization in finding people in natural disasters. The samples can be collected by the government during census or other scenarios and store in the database and that can be used in various criminal and forensic applications. In future, in case of mutation in the genomic sequence, by examining the variations in the gene sequence the various diseases affected in an individual is found out.

REFERENCES

[1] Ferri, G., *et al.*, "Species identification through DNA "barcodes"," *Genetic Testing and Molecular Biomarkers*, vol/issue: 13(3), pp. 421-426, 2009.
[2] Wilson J. J., "DNA barcodes for insects," *DNA barcodes: Methods and protocols*, pp. 17-46, 2012.
[3] Ward, R. D., *et al.*, "The campaign to DNA barcode all fishes, FISH-BOL," *Journal of fish biology*, vol/issue: 74(2), pp. 329-356, 2009.
[4] Cywinska A., *et al.*, "Identifying Canadian mosquito species through DNA barcodes," *Medical and veterinary entomology*, vol/issue: 20(4), pp. 413-24, 2006.
[5] Barrett R. D. and Hebert P. D., "Identifying spiders through DNA barcodes," *Canadian Journal of Zoology*, vol/issue: 83(3), pp. 481-91, 2005.
[6] Hebert P. D., *et al.*, "Identification of birds through DNA barcodes," *PLoS Biol.*, vol/issue: 2(10), pp. e312, 2004.
[7] Hebert P. D., *et al.*, "Biological identifications through DNA barcodes," *Proceedings of the Royal Society of London B: Biological Sciences*, vol/issue: 270(1512), pp. 313-21, 2003.
[8] Witt J. D., *et al.*, "DNA barcoding reveals extraordinary cryptic diversity in an amphipod genus: implications for desert spring conservation," *Molecular Ecology*, vol/issue: 15(10), pp. 3073-82, 2006.
[9] Ball S. L. and Armstrong K. F., "DNA barcodes for insect pest identification: a test case with tussock moths (Lepidoptera: Lymantriidae)," *Canadian Journal of Forest Research*, vol/issue: 36(2), pp. 337-50, 2006.
[10] Zokaee S. and Faez K., "Human identification based on ECG and palmprint," *International Journal of Electrical and Computer Engineering*, vol/issue: 2(2), pp. 261, 2012.
[11] Kim Y., *et al.*, "The nucleotide: DNA sequencing and its clinical application," *Journal of oral and maxillofacial surgery*, vol/issue: 60(8), pp. 924-30, 2002.
[12] L. Wang and C. A. Alexander, "Applications of Automated Identification Technology in EHR/EMR," *International Journal of Public Health Science (IJPHS)*, vol/issue: 2(3), pp. 109-122, 2013.
[13] E. S. Orabi, *et al.*, "DNA fingerprint using smith waterman algorithm by grid computing," in *Informatics and Systems (INFOS), 2014 9th International Conference on (ppPDC-74). IEEE*, 2014.
[14] S. A. Shehab, *et al.*, "Fast dynamic algorithm for sequence alignment based on bioinformatics," *International Journal of Computer Applications*, vol/issue: 37(7), pp. 54-61, 2012.
[15] P. Adhitama, *et al.*, "Lexicon-Driven Word Recognition Based on Levenshtein Distance," *International Journal of Software Engineering and Its Applications*, vol/issue: 8(2), pp. 11-20, 2014.
[16] Benson G., "Tandem repeats finder: a program to analyze DNA sequences," *Nucleic acids research*, vol/issue: 27(2), pp. 573, 1999.
[17] Kolpakov R., *et al.*, "Mreps: efficient and flexible detection of tandem repeats in DNA," *Nucleic acids research*, vol/issue: 31(13), pp. 3672-8, 2003.
[18] Ruitberg C. M., *et al.*, "STRBase: a short tandem repeat DNA database for the human identity testing community," *Nucleic Acids Research*, vol/issue: 29(1), pp. 320-2, 2001.

BIOGRAPHIES OF AUTHORS

Likhitha C P was born in Coorg-India in 1992. She received the BCA degree in Computer Science from the Amrita Vishwa Vidyapeetham (Amrita University), Mysuru Campus, India, in 2014. Currently, she is persuing her MCA degree in Computer Science from the Amrita Vishwa Vidyapeetham (Amrita University), Mysuru Campus, India. Her research interests include Bioinformatics and Image Processing.

Ninitha P was born in Mysuru-India in 1993. She received the BCA degree in Computer Science from the Amrita Vishwa Vidyapeetham (Amrita University), Mysuru Campus, India, in 2014. Currently, she is persuing her MCA degree in Computer Science from the Amrita Vishwa Vidyapeetham (Amrita University), Mysuru Campus, India. Her research interests include Bioinformatics and Software Engineering.

Kanchana V was born in Mysuru-India in 1979. She received the B.Sc. degree in PMCS from University of Mysuru, the MCA degree in Computer Science from VTU, and the MTech degree in IT from Karnataka State Open University. She has more than 11 years of academic experience. Currently, she is working as Assistant Professor in Department of Computer Science, Amrita Vishwa Vidyapeetham (Amrita University), Mysuru Campus, India. Her research areas include Bioinformatics, MIS, ERP and Software Engineering.