Proceedings of the Pakistan Academy of Sciences 49 (4): 251–257 (2012)
Copyright © Pakistan Academy of Sciences
ISSN: 0377 - 2969 print / 2306 - 1448 online

Pakistan Academy of Sciences

Research Article

MATLAB-based Sequence Analysis of *muRdr1H*, a Functionally Characterized Resistance Gene of Roses

Aneela Yasmin[1], Akhtar A. Jalbani[2]* and Abdul Samad Chandio[3]

[1]Department of Biotechnology, Faculty of Crop Production,
Sindh Agriculture University, Tandojam-70060, Pakistan
[2]Information Technology Center, Sindh Agriculture University,
Tandojam-70060, Pakistan
[3]Department of Irrigation and Drainage, Faculty of Agricultural Engineering,
Sindh Agriculture University, Tandojam-70060, Pakistan

Abstract: One of the practical applications of Bioinformatics is sequence analysis through sequence alignment of newly identified genes and previously available sequences to determine the similarity between the biological sequences and elucidate the functionality of genes with confidence. Although there are many softwares available for the biological sequence alignment; in this paper some basic alignment algorithms are discussed and implemented in MATLAB Bioinformatics tool box. Two sequences analyzed by this tool box were *muRdr1H* and resistance protein of *Populus trichocarpa* (ACCESSION XP_002329162). *muRdr1H* is a functionally characterized member of *Rdr1* resistance gene family of roses, active against black spot. The simulated result shows that the MATLAB tool is a comprehensive tool for finding best possible local and global sequence alignment. Our proposed work provides useful application of MATLAB which can help in interpretation and visualization of the data in molecular biology.

Keywords: Sequence alignment, MATLAB, Bioinformatics, protein sequence

1. INTRODUCTION

Computer science in combination with biological science is a relatively new multidisciplinary area of science, known as Bioinformatics. In this research area, Informatics involves the technology that uses computer for different purposes; for example, storage of biological data and for performing various techniques on the data for efficient retrieval, manipulation and distribution. As the data contains biological information, hence the distribution of information is related to biological aspects, such as DNA, RNA and proteins [1]. In this domain, research focuses on the usage of computers because most of the tasks in genomics data analysis are highly repetitive. Moreover, the common activities performed in Bioinformatics research include mapping and analyzing or aligning DNA and protein sequences, comparing and creating 3-D models of protein structures, gene finding, genome assembly, drug design and discovery.

In computational biology the main aspects of using Informatics is to develop innovative algorithms to compute biological sequence-related problems. In handling biological sequences, sequence alignment is a key process of Bioinformatics and computational biology. During alignment, sequences are compared by identifying similar patterns and establishing residue-residue correspondence among related sequences [2, 3]. Hence sequence alignment is a way of arranging primary sequences of DNA, RNA and proteins to detect similar regions that may be consequences of

Received, May 2012; Accepted, November 2012

*Corresponding author, Akhtar A. Jalbani; Email: akjalbani@sau.edu.pk functional, structural or evolutionary relationship between the sequences. The resulting alignment produces revise transcript of mismatches, i.e., insertions and deletions, where mismatches can be inferred as point mutations. As a result, we can infer how sequences with the identical origin would deviate from one another.

Nowadays, new biological sequences are being generated at an exponential rate; hence sequence comparison is widely used in biological research to identify new proteins and can search for existing protein for drug or diseases discovery. In any genome project, newly determined sequences are first compared with those which are already present in the genomic databases, such as NCBI, in order to discover similarities. As a result of comparison, one or more sequence alignments can be produced. Similarity score is one factor that can be associated with the sequence alignment. If new sequences are identified then that sequenceMATLABs are added to the biological databases, like NCBI, ENTREZ (which integrates GenBank) [4]. These databanks are remotely accessible. Researchers take full advantages of these databases to query and compare their sequences using different Bioinformatics tools.

Thus, the sequence alignment is the first step to expose the structural or functional importance of unknown sequences. Here in this paper, two very important types of DNA alignments are discussed namely Global and Local alignment in MATLAB. MATLAB is a powerful tool for the modeling and simulation of various domains. It contains different toolbox for the different domain for example communication systems, image compression, Bioinformatics, etc. MATLAB provides tight integration of compiler and its simulating environment. In this paper, we focused on the Bioinformatics tool box, which contains different solutions for the biological research; sequence alignment is one of them. In local alignment, the assumption is made based on that the two sequences are not similar over entire length [5]. Hence it only identifies the local regions with highest level of similarities and the aligned two sequences show region based alignment without considering the alignment of the rest of the regions. It is possible that the two sequences to be aligned can be of different lengths [6, 7] whereas global alignment uses dynamic programming to obtain global alignment between the two selected sequences. These two algorithms were applied in MATLAB.

The two sequences selected to analyze in MATLAB were *muRdr1H* and TIR-NBS-LRR-resistance protein of *Populus trichocarpa* (ACCESSION XP_002329162). *muRdr1H* is a member of *Rdr1* resistance gene family of roses active against black spot, a fungal disease. This gene is functionally characterized and was identified as active resistance genes against Dort E4, race 6 of *diplocarpon rosae* [8, 9] whereas TIR-NBS-LRR-resistance protein of *Populus trichocarpa* shares the highest identity (41%) to *muRdr1H* protein. It is important to state that the TIR-NBS-LRR-resistance protein of *Populus trichocarpa* was not functionally characterized.

2. MATERIALS AND METHODS

To identify the homologues of *muRdr1H* gene BLASTx searches were carried out against the GeneBank non-redundant database (http://blast.ncbi.nlm.nih.gov). The selected two sequences were tested using MATLAB Bioinformatics software tools (www.mathworks.com). Those two sequences are assumed to be identical in length. The sequence alignment is carried out from beginning to the end of both sequences to find the best possible sequence alignment. The two types of alignment methods considered for the practical assignment were local alignment and global alignment.

2.1 Algorithms for the DNA Sequence Alignment

DNA sequence strings consist of four alphabet letters (A, C, G, T) called nucleotides. The length of sequence is variable; hence algorithm should produce high quality sequence alignment from these four letters. In local alignment, we need to identify the isolated regions of high similarity from the entire DNA sequence that makes better choice in some situation for this type of alignment method but it is more complex in general [10].

2.1.1. Local Alignment

The local alignment is performed through very well-known algorithm known as Smith-Waterman algorithm. It is used for determining identical regions between two nucleotides or proteins. The algorithm does not look for total sequence but it compares segments of all possible lengths and optimizes the similarity measures. Following steps should be considered for Smith-Waterman algorithm:

Step-1: Fill all dynamic matrix

Step-2: To find optimal local alignment length, find max value and trace of max value for that patch which leads to the max value or score.

The Smith-Waterman algorithm compares two DNA sequences based individual pair-wise comparison between the characteristics.

2.1.2. Global Alignment

Global alignment method uses dynamic programming, for this Needleman-Wunsch algorithm is the best algorithm for this type of sequence alignment. From two sequences, this algorithm helps to identify the global alignment. This uses all elements of the two sequences for the alignment procedure. This is also called end-to-end alignment.

All elements are considered in global sequence alignment method, therefore the scoring matrix will also become m*n (where m is the longer sequence and n is the shorter sequence). The optimal score can be calculated at each matrix position by adding current match score to last scored position and subtracting the gap penalties. Hence each matrix position may have +ve or –ve or even zero value.

2.2. Sequence Alignment with MATLAB

The algorithms discussed above can easily be applied in MATLAB to get most efficient or the best possible sequence alignment for nucleotide or protein. MATLAB contains different built-in functions to access the already stored sequencing data on the gene databanks. Using MATLAB we can apply global and local alignment method to find metrics scores. In MATLAB, sequences were entered by accession numbers of the sequences, sequences were retrieved in its Open Reading Frames (ORF). After bringing the information in MATLAB environment, we applied algorithms for comparison of sequences using global and local alignment with the score that determines the degree of similarity. At the end, Monte Carlo Techniques were applied in MATLAB environment to assess the significance of alignment. For this random sequences were generated and their scores were plotted as bars; then using the evfit function from the statistics toolbox, the parameters of bar distribution were estimated and the probability density function of the estimated distribution was plotted in red line (Fig. 1).

3. RESULTS AND DISCUSSION

3.1. Selection of Sequences to Study Alignment Algorithms

According to BLASTx searches *muRdr1H* protein shares the maximum, i.e., 41%, identity to TIR-NBS-LRR-resistance proteins of *Populus trichocarpa* (ACCESSION XP_002329162), both genes belong to the same class of TIR-NBS-LRR resistance genes, followed by 39-44% to hypothetical proteins of *Vitis vinifera*, 40% to TIR of *Medicago truncatula* (ACCESSION ABD28703), 40% to *CMR1* of *Phaseolus vulgaris* (ACCESSION ABH07384) and 39% identity to *N*-like protein of *N. tabacum* (ACCESSION BAF95888) for resistance to the Tobacco Mosaic Virus. Out of these sequences TIR-NBS-LRR-resistance proteins of *Populus trichocarpa* was selected for the alignment with *muRdr1H* using MATLAB to demonstrate the utility of this software for searching differences in their protein sequences.

3.2. Open Reading Frame of Selected Sequences in MATLAB Environment

The protein sequences of both genes were in text format. We used the same sequence by using seqshoworfs MATLAB method to open it in open reading frame work reader (ORF) as shown in Fig. 2.

3.3. Dot Plot based Comparison in MATLAB Environment

The dot plot representation of selected sequences is shown in Fig. 3. This basic sequence alignment method compares two sequences in graphical method by comparing two dimensional matrix; two sequences are written in vertical and horizontal

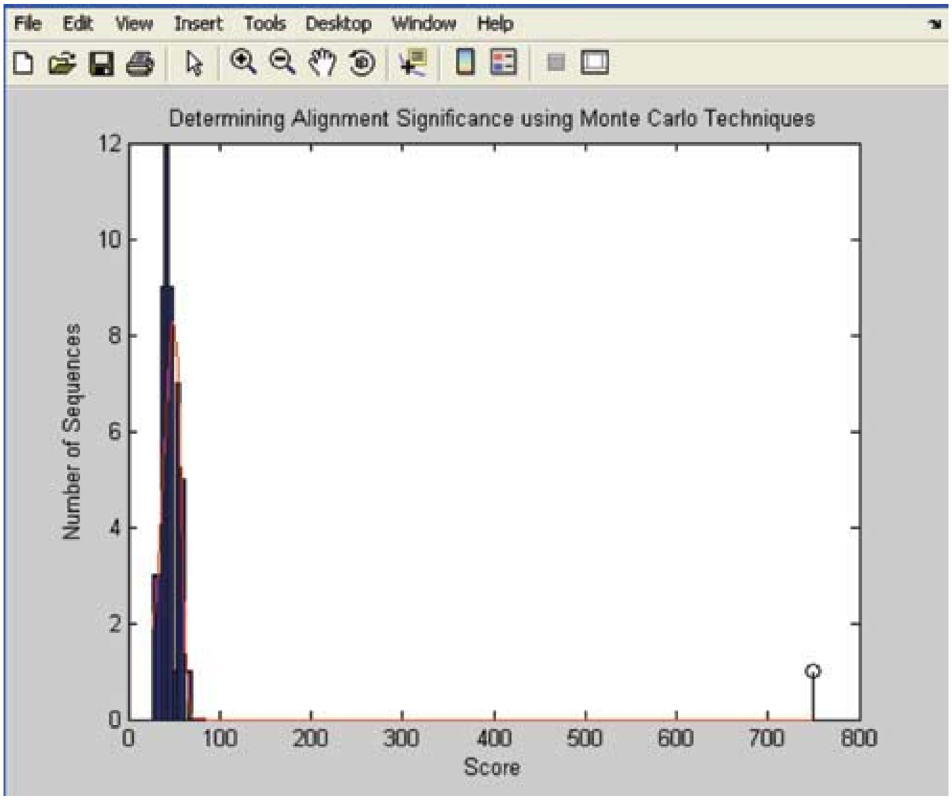
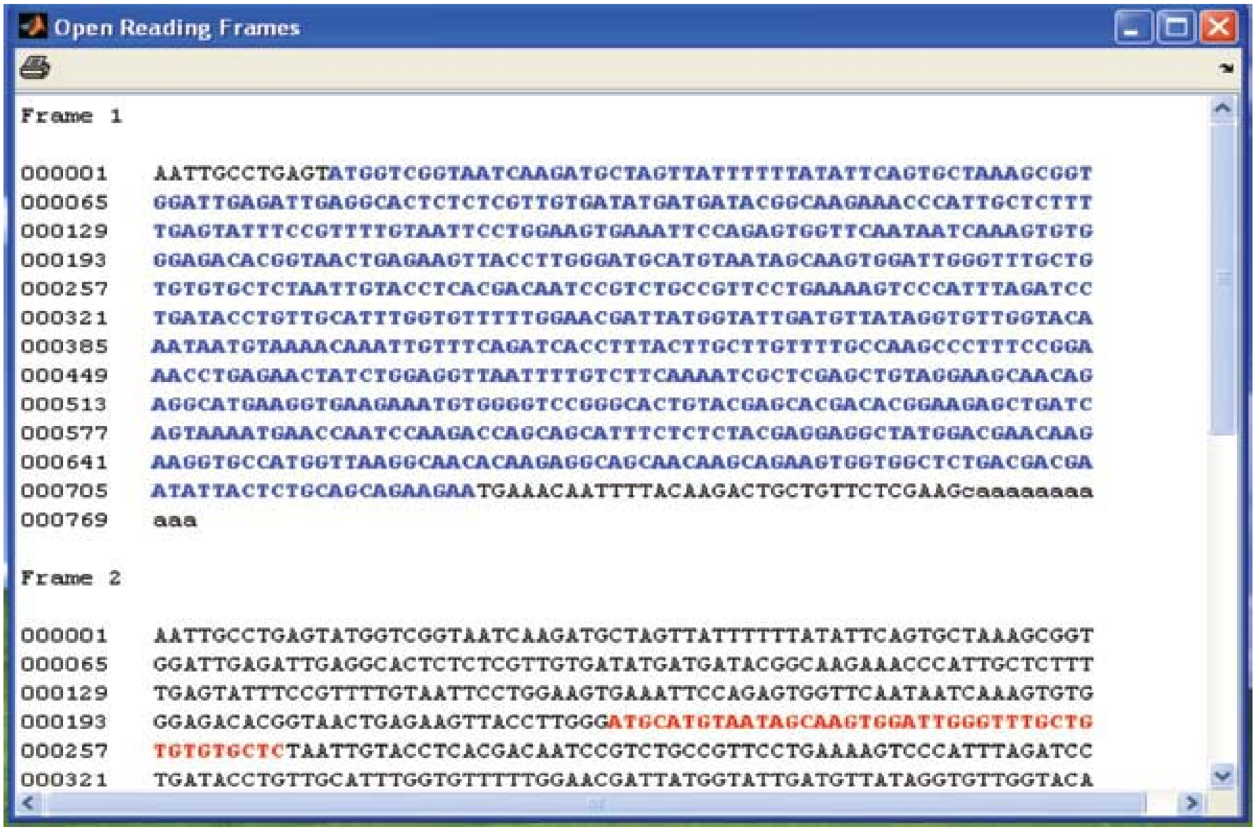


Fig. 1. Significance of Alignments between *muRdr1H* and TIR-NBS-LRR resistance protein of *Populus trichocarpa* are assessed using Monte Carlo Techniques.



```
Frame 1

000001    AATTGCCTGAGTATGGTCGGTAATCAAGATGCTAGTTATTTTTTATATTCAGTGCTAAAGCGGT
000065    GGATTGAGATTGAGGCACTCTCTCGTTGTGATATGATGATACGGCAAGAAACCCATTGCTCTTT
000129    TGAGTATTTCCGTTTTGTAATTCCTGGAAGTGAAATTCCAGAGTGGTTCAATAATCAAAGTGTG
000193    GGAGACACGGTAACTGAGAAGTTACCTTGGGATGCATGTAATAGCAAGTGGATTGGGTTTGCTG
000257    TGTGTGCTCTAATTGTACCTCACGACAATCCGTCTGCCGTTCCTGAAAAGTCCCATTTAGATCC
000321    TGATACCTGTTGCATTTGGTGTTTTTGGAACGATTATGGTATTGATGTTATAGGTGTTGGTACA
000385    AATAATGTAAAACAAATTGTTTCAGATCACCTTTACTTGCTTGTTTTGCCAAGCCCTTTCCGGA
000449    AACCTGAGAACTATCTGGAGGTTAATTTTGTCTTCAAAATCGCTCGAGCTGTAGGAAGCAACAG
000513    AGGCATGAAGGTGAAGAAATGTGGGGTCCGGGCACTGTACGAGCACGACACGGAAGAGCTGATC
000577    AGTAAAATGAACCAATCCAAGACCAGCAGCATTTCTCTCTACGAGGAGGCTATGGACGAACAAG
000641    AAGGTGCCATGGTTAAGGCAACACAAGAGGCAGCAACAAGCAGAAGTGGTGGCTCTGACGACGA
000705    ATATTACTCTGCAGCAGAAGAATGAAACAATTTTACAAGACTGCTGTTCTCGAAGcaaaaaaaa
000769    aaa

Frame 2

000001    AATTGCCTGAGTATGGTCGGTAATCAAGATGCTAGTTATTTTTTATATTCAGTGCTAAAGCGGT
000065    GGATTGAGATTGAGGCACTCTCTCGTTGTGATATGATGATACGGCAAGAAACCCATTGCTCTTT
000129    TGAGTATTTCCGTTTTGTAATTCCTGGAAGTGAAATTCCAGAGTGGTTCAATAATCAAAGTGTG
000193    GGAGACACGGTAACTGAGAAGTTACCTTGGGATGCATGTAATAGCAAGTGGATTGGGTTTGCTG
000257    TGTGTGCTCTAATTGTACCTCACGACAATCCGTCTGCCGTTCCTGAAAAGTCCCATTTAGATCC
000321    TGATACCTGTTGCATTTGGTGTTTTTGGAACGATTATGGTATTGATGTTATAGGTGTTGGTACA
```

Fig. 2. Open Reading Framework (ORF) options for *muRdr1H* Sequence in MATLAB.

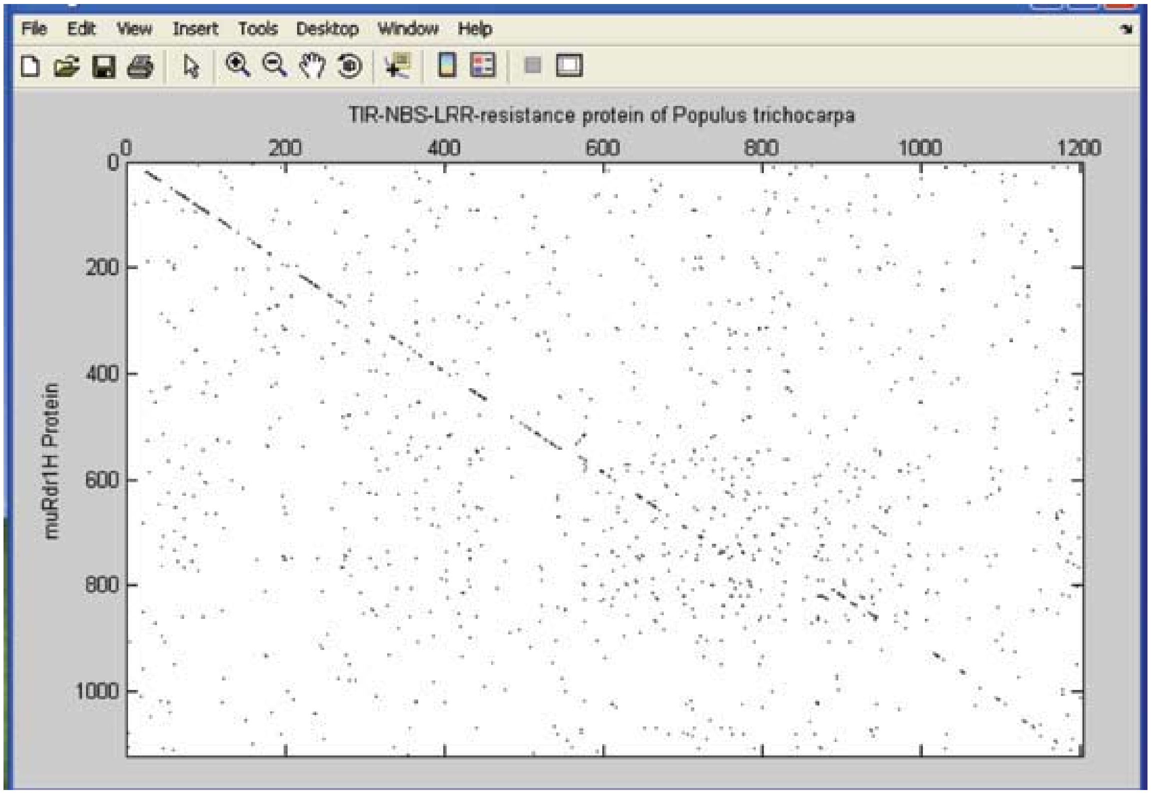
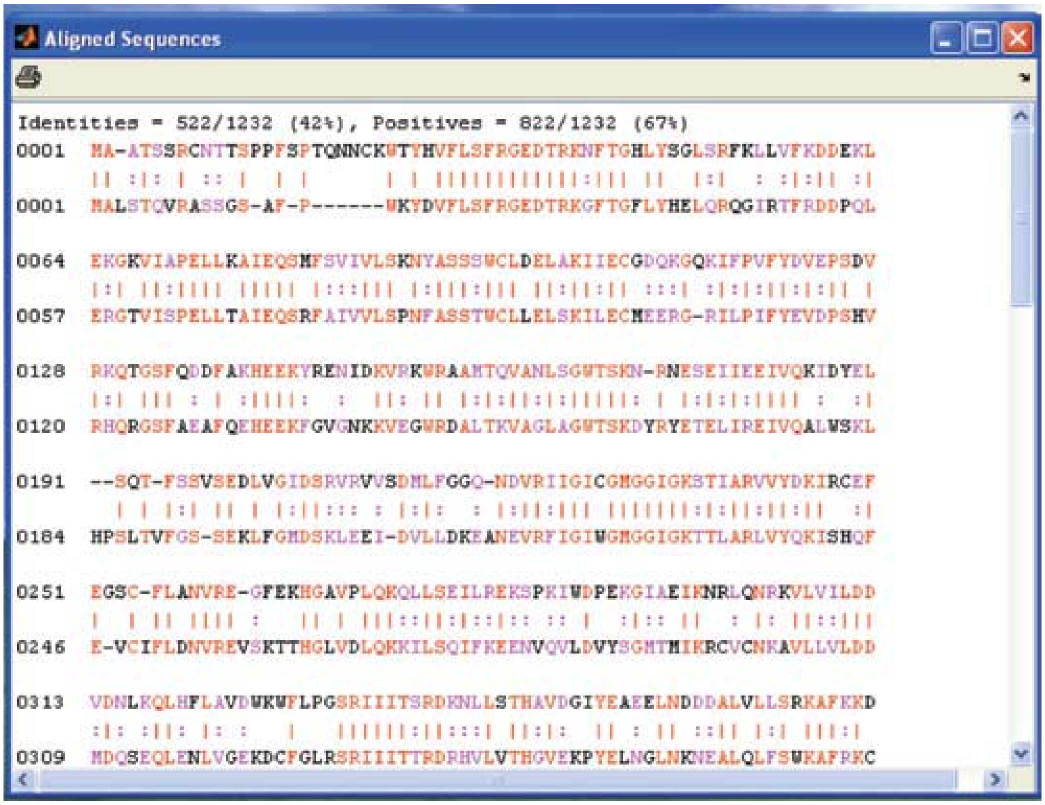
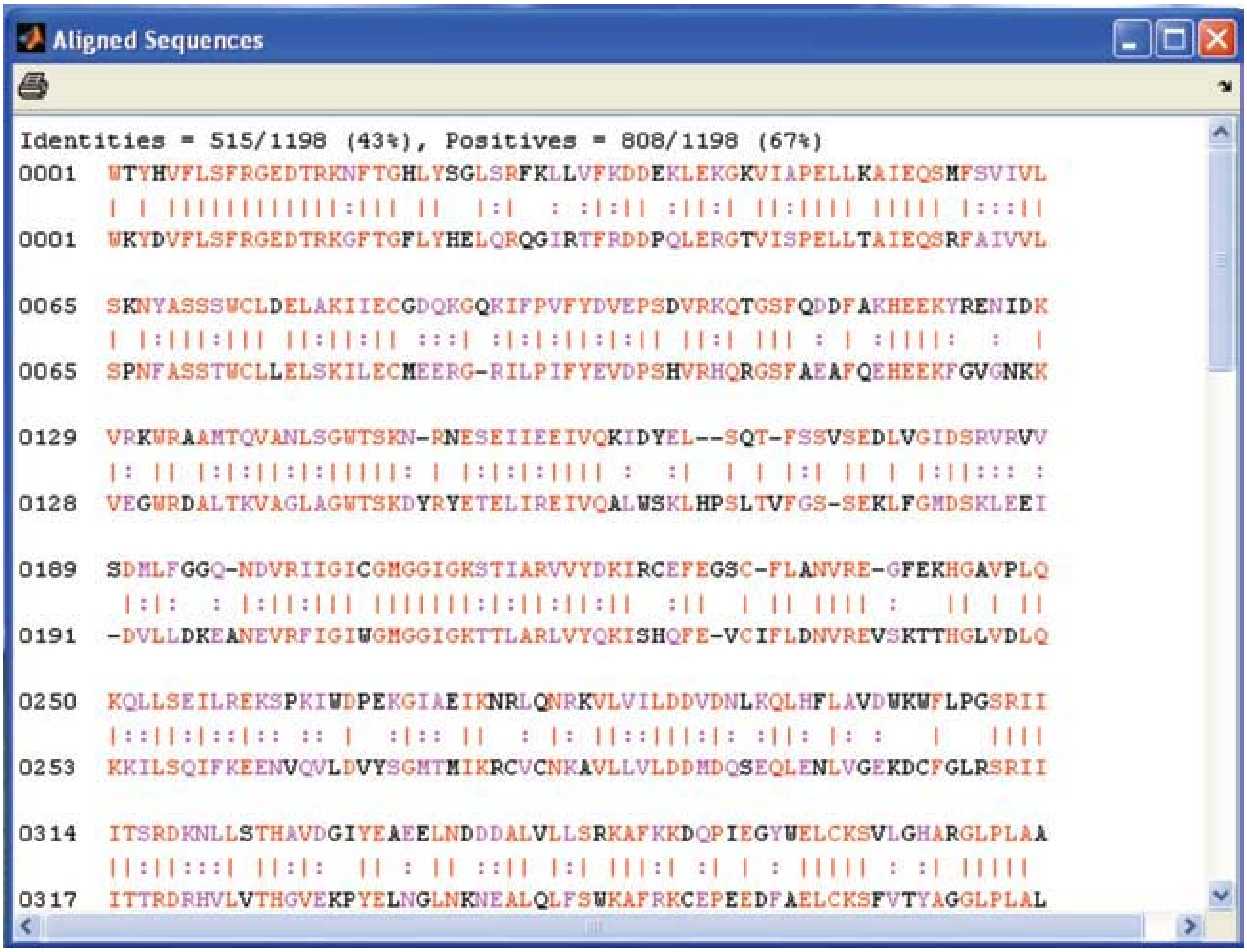


Fig. 3. Dot plot of *muRdr1H* and TIR-NBS-LRR resistance protein of *Populus trichocarpa in MATLAB* environment.



```
Identities = 522/1232 (42%), Positives = 822/1232 (67%)
0001  MA-ATSSRCNTTSPPFSPTQNNCKWTYHVFLSFRGEDTRKNFTGHLYSGLSRFKLLVFKDDEKL
      || :|: | :: |  | |        | | ||||||||||||:||| ||  |:|  : :|:|| :|
0001  MALSTQVRASSGS-AF-P-------WKYDVFLSFRGEDTRKGFTGFLYHELQRQGIRTFRDDPQL

0064  EKGKVIAPELLKAIEQSMFSVIVLSKNYASSSWCLDELAKIIECGDQKGQKIFPVFYDVEPSDV
      |:| ||:|||| ||||| |:::||| |:|||:||| ||:||:|| :::| :|:|:||:|:|| |
0057  ERGTVISPELLTAIEQSRFAIVVLSPNFASSTWCLLELSKILECMEERG-RILPIFYEVDPSHV

0128  RKQTGSFQDDFAKHEEKYRENIDKVRKWRAAMTQVANLSGWTSKN-RNESEIIEEIVQKIDYEL
      |:| ||| : | :||||:  :  ||: || |:|:||:|:|||||: | |:|:|:|||| :  :|
0120  RHQRGSFAEAFQEHEEKFGVGNKKVEGWRDALTKVAGLAGWTSKDYRYETELIREIVQALWSKL

0191  --SQT-FSSVSEDLVGIDSRVRVVSDMLFGGQ-NDVRIIGICGMGGIGKSTIARVVYDKIRCEF
        | | |:| || | |:||::: : |:|:  : |:||:||| ||||||||:|:||:||:||  :|
0184  HPSLTVFGS-SEKLFGMDSKLEEI-DVLLDKEANEVRFIGIWGMGGIGKTTLARLVYQKISHQF

0251  EGSC-FLANVRE-GFEKHGAVPLQKQLLSEILREKSPKIWDPEKGIAEIKNRLQNRKVLVILDD
      |  | || |||| :   || | |||::||:|::|:: :: |  :|:: ||  : |: |||::|||
0246  E-VCIFLDNVREVSKTTHGLVDLQKKILSQIFKEENVQVLDVYSGMTMIKRCVCNKAVLLVLDD

0313  VDNLKQLHFLAVDWKWFLPGSRIIITSRDKNLLSTHAVDGIYEAEELNDDDALVLLSRKAFKKD
      :|: :||: |: :    |    ||||||:||:::| ||:|:  || : || ::|| |:| ||||:|
0309  MDQSEQLENLVGEKDCFGLRSRIIITTRDRHVLVTHGVEKPYELNGLNKNEALQLFSWKAFRKC
```

Fig. 4. Global Alignment of *muRdr1H* and TIR-NBS-LRR resistance protein of *Populus trichocarpa.*



```
Identities = 515/1198 (43%), Positives = 808/1198 (67%)
0001   WTYHVFLSFRGEDTRKNFTGHLYSGLSRFKLLVFKDDEKLEKGKVIAPELLKAIEQSMFSVIVL
       | | ||||||||||||:||| ||  |:|  : :|:|| :||:| ||:|||| ||||| |:::||
0001   WKYDVFLSFRGEDTRKGFTGFLYHELQRQGIRTFRDDPQLERGTVISPELLTAIEQSRFAIVVL

0065   SKNYASSSWCLDELAKIIECGDQKGQKIFPVFYDVEPSDVRKQTGSFQDDFAKHEEKYRENIDK
       | |:|||:||| ||:||:|| :::| :|:|:||:|:|| ||:| ||| : | :||||:  :  |
0065   SPNFASSTWCLLELSKILECMEERG-RILPIFYEVDPSHVRHQRGSFAEAFQEHEEKFGVGNKK

0129   VRKWRAAMTQVANLSGWTSKN-RNESEIIEEIVQKIDYEL--SQT-FSSVSEDLVGIDSRVRVV
       |: || |:|:||:|:|||||: | |:|:|:|||| :  :|  | | |:| || | |:||::: :
0128   VEGWRDALTKVAGLAGWTSKDYRYETELIREIVQALWSKLHPSLTVFGS-SEKLFGMDSKLEEI

0189   SDMLFGGQ-NDVRIIGICGMGGIGKSTIARVVYDKIRCEFEGSC-FLANVRE-GFEKHGAVPLQ
        |:|:  : |:||:||| |||||||:|:||:||:||  :||  | || |||| :   || | ||
0191   -DVLLDKEANEVRFIGIWGMGGIGKTTLARLVYQKISHQFE-VCIFLDNVREVSKTTHGLVDLQ

0250   KQLLSEILREKSPKIWDPEKGIAEIKNRLQNRKVLVILDDVDNLKQLHFLAVDWKWFLPGSRII
       |::||:|::|:: :: |  :|:: ||  : |: ||::|||:|: :||: |: :   |   ||||
0253   KKILSQIFKEENVQVLDVYSGMTMIKRCVCNKAVLLVLDDMDQSEQLENLVGEKDCFGLRSRII

0314   ITSRDKNLLSTHAVDGIYEAEELNDDDALVLLSRKAFKKDQPIEGYWELCKSVLGHARGLPLAA
       ||:||:::| ||:|:  || : || ::|| |:| |||:| :| | : ||||| : :| |||||
0317   ITTRDRHVLVTHGVEKPYELNGLNKNEALQLFSWKAFRKCEPEEDFAELCKSFVTYAGGLPLAL
```

Fig. 5. Local Alignment of of *muRdr1H* and TIR-NBS-LRR resistance protein of *Populus trichocarpa.*

axes of the matrix and compared [10]. The selected sequences exhibited substantial regions of similarity (Fig. 3). Many dots are lined up in a diagonal line revealing some sequence alignment that will be confirmed and regions will be identified by local and global alignment. It is obvious from the line that TIR and NBS region of protein has more similarity as compared to the end region that represents LRR repeats. This result reflected the initial identity of both proteins, which was 41% as described above.

3.4. Global and Local Sequence Alignments

Both sequences are compared using global and local alignment methods. For global alignment we used Needle-Wunsch algorithm in the MATLAB and the resultant aligned global sequence is shown in Fig 3. The scoring of global alignment is 749.33 with 522/ 1232 (42%) identities and 822/ 1232 (67%) positives. The same sequences were used for local alignment. In which we used Smith-Waterman algorithm, where pair wise comparison is shown in figure 4. The matrix score of local alignment is 766.66 with 43% identities and 67% positives. Sequences with good similarity did not show much difference between local and global alignments. Nevertheless, local alignment is better to find conserved regions of a gene.

The number of identities and positives in global and local alignments are shown in Fig. 4 and Fig. 5, respectively. Actually this did not represent the significant alignment. So we applied Monte Carlo Techniques in MATLAB environment to assess the significance of an alignment as described in methods. The random sequences were generated in MATLAB and their scores were plotted as bars (Fig 1). The evfit function from the statistics toolbox gave the scores of 49.28 and 9.5. The probability density function of the estimated distribution was plotted which clearly showed that global alignment was significant (Fig. 5).

In this paper, global and local alignment algorithms were developed and simulated using MATLAB. For global alignment, Needleman-Wunsch algorithm and for local alignment Smith-Waterman algorithm were used in MATLAB. This demonstration of data mining, visualization and of course interpretation in MATLAB can facilitate the sequence data handling in molecular biology.

4. REFERENCES

1. Mathkour, H. & M. Ahmad. A comprehensive survey on genome sequence analysis. In: *IEEE International Conference on Bioinformatics and Biomedical Technology*, p. 14-18 (2010).
2. Benkrid, K., Y. Liu & A. Benkrid. A highly parameterized and efficient FPGA-based skeleton for pairwise biological sequence alignment. *IEEE Transactions on Very Large Scale Integration Systems* 17(4): 561-570 (2009).
3. Boukerche, A., J. M. Correa, A. Cristina M.A de Melo & R. P. Jacob. A hardware accelerator for the fast retrieval of DIALIGN biological sequence alignments in linear space. *IEEE Transactions on Computers* 59(6): 808-821 (2010).
4. Hong, C. & A. Tewfic, heuristic reusable dynamic programming: Efficient updates of local sequence alignment. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 6(4): 570-562 (2009).
5. Nguyen, V., A. Cornu & D. Lavenier. Implementing protein seed-based comparison algorithm on the SGI RASC-100 platform. In: *IEEE Symposium on Parallel and Distributed Processing*. p. 1-7 (2009).
6. Bandyopadhayay, S. & R. Mitra. A parallel pairwise local sequence alignment algorithm. *IEEE Transactions on Nano Bioscience* 8(2): 139-146 (2009).
7. Nahar, N., M. Popstova & J. Gogarten. GPX: A tool for the exploration and visualization of genome evolution. In: *7th IEEE International Conference on Bioinformatics and Bioengineering*, p. 1338-1342 (2007).
8. Yasmin, A. *Identification and Molecular Characterization of Rdr1 Resistance Gene from Roses*. PhD thesis, University of Hannover, Germany (2010).
9. Terefe-Ayana, D., A. Yasmin, T. L. Le, H. Kaufmann, A. Biber, A. Kuehr, M. Linde & T. Debener. Mining disease resistance genes in roses: functional and molecular characterization of the *Rdr1* locus. *Frontiors of Plant Science* 2:35, doi: 10.3389/fpls.2011.00035 (2011).
10. Mai, S. M., M. Hamdy, M. Aboelfotoh & Y. M. Kadah. BIOINFTool: Bioinformatics and sequence data analysis in molecular biology using MATLAB. In: *Cairo International Biomedical Engineering Conference*, p. 1-9 (2006).