Cihan University-Erbil Scientific Journal (CUESJ)

Research Article

Measuring the Score Matching of the Pairwise Deoxyribonucleic Acid Sequencing using Neuro-Fuzzy

Safa A. Hameed*, Raed I. Hamed

Department of Computer Science, College of Engineering and Science, University of Bayan, Erbil, Iraq

ABSTRACT

The proposed model for getting the score matching of the deoxyribonucleic acid (DNA) sequence is introduced; the Neuro-Fuzzy procedure is the strategy actualized in this paper; it is used the collection of biological information of the DNA sequence performing with global and local calculations so as to advance the ideal arrangement; we utilize the pairwise DNA sequence alignment to gauge the score of the likeness, which depend on information gathering from the pairwise DNA series to be embedded into the implicit framework; an adaptive neuro-fuzzy inference system model is reasonable for foreseeing the matching score through the preparation and testing in neural system and the induction fuzzy system in fuzzy logic that accomplishes the outcome in elite execution.

Keywords: Component, dynamic programming, matching, neuro-fuzzy, sequence alignment

INTRODUCTION

Deoxyribonucleic acid (DNA) sequence matching is an essential area and more approaching nearby in computational biological data.[1] DNA sequence analysis is an imperative exploration topic in bioinformatics. Assessing the similarity between sequences, this is important for sequence analysis, because similarity proves congruence.[2] The DNA atom contains biological, physical, and chemical data; it has turned out to be essential to examine DNA sequences statistically.[3] String matching is a strategy to find a design from the predefined info string.[4] Similarities between DNA sequences may emerge due to the functional, structural, or transformative relationship among them.[5] Sequence alignment of two biological sequences may be called pairwise sequence alignment, also in the event, more than two sequences are involved; it may be called multiple sequence alignment.[6] The dynamic programming is the method to implement the DNA alignment using the Needleman–Wunsch[7] and Smith–Waterman algorithms.[8]

Here, in this article, we use the pairwise sequence alignment in a global and local algorithms and examined the measure of the matching based on the collected data for DNA alignment. The Neuro-Fuzzy model[9] is used in the Matlab tool that implemented by the data set files of measure score matching of DNA sequences that deal with the set of biological data. This tool is efficient and fast to evaluate the scoring measure of matching the DNA sequences.

LITERATURE REVIEW

The study of a biological sequence has been growing exponentially, while the applications of the sequence alignment cover the wide range in bioinformatics. The previous research work has been studied to provide new Algorithms with the main purpose for the requirements of matching sequences; the techniques have been used all the latest with providing fast and efficient sequence alignment algorithms. Bhukya and Somayajulu[1] suggested a new pattern for matching technique defined as exact multiple pattern-matching algorithms that utilize DNA sequence. The current method is used to avoid unneeded comparisons in the DNA sequence. Gill and Singh[6] proposed a multiple sequence alignment algorithm which performs fuzzy logic to measure the similarity of sequences based on the fuzzy parameters. Nasser *et al*.[10] suggested the fuzzy logic model for approximate matching of DNA subsequences. Kim *et al*.[11] suggested a DNA sequence alignment, which uses quality information and a fuzzy inference implementation developed based on the features of DNA parts and a fuzzy logic system. Chai *et al*.[12] explained how to perform pairwise sequence alignments utilizing the biostrings bundle using the pairwise alignment function. Hameed and Hamed[13] discussed how to

Corresponding Author:
Safa A. Hameed, Department of Computer Science, College of Engineering and Science, University of Bayan, Erbil, Iraq.
E-mail:safa.hamid@bnu.edu.iq

Received: Mar 21, 2019
Accepted: Apr 24, 2019
Published: Aug 20, 2019

DOI: 10.24086/cuesj.v3n2y2019.pp37-41

implement the pairwise alignment technique to get the score of similarity for a pair of characters. In our work, we use the Neuro-Fuzzy model that utilizes the biological dataset files for matching DNA and measures the score of matching the DNA sequences with global and local alignment.

SEQUENCE ALIGNMENT

DNA matching is a significant venture in the sequence alignment. Since sequence alignment is a discretionary matching process, there is a need for better algorithms.[10] DNA sequence alignment algorithms over computational biological science have been enhanced eventually by different techniques:[11] the (Needleman-Wunsch) global, the (Smith-Waterman) local, and (ends-free) cover pairwise sequence alignment issues.[12] Pairwise alignment is a technique for scoring the similarity of a pair of characters. It decides the correspondences between the substrings in the sequences like the similarity score is amplified.[13] For it is a large portion basic form, known as pairwise sequence alignment, we provided for two sequences A and B and discover their best alignment (either global or local).[12] Aligned sequences represented as rows in a grid. Gaps ("−") need aid embedded between the characters with the goal.[6] The ways we use it to perform the alignment are global and local alignment, these algorithms uses the proposed matrix to measure the similarity of bases in the two sequences. For the Needleman–Wunsch algorithm, a scoring matrix is ascertained for those two provided for sequences A and B, by setting one sequence along column side, furthermore on the turn sequence side. It is additionally frequently referred as optimal matching algorithm and the global alignment technique.[7] The Smith–Waterman algorithm, which is the method used to perform the local sequence alignment, local alignment algorithms find the sections of the highest similarity between two sequences and create the alignment to abroad from there, that is, identify the most similar portion comparable subregion imparted between two successions.[8]

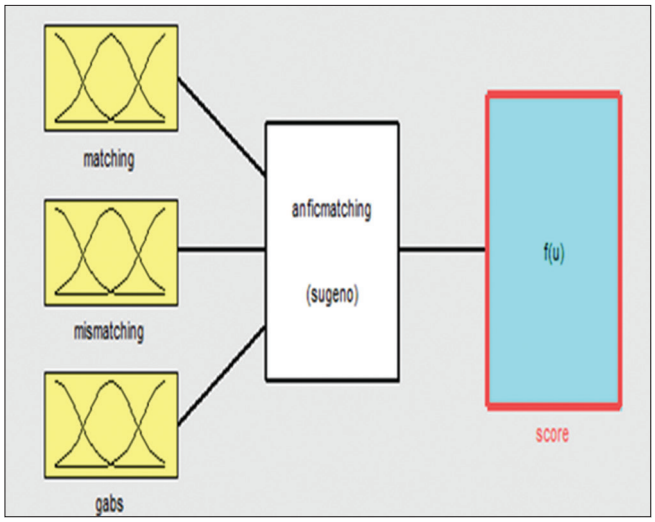

Figure 1: The three input variables and the output

THE PROPOSED METHOD

We use the Nero-Fuzzy technique in Matlab tool. The Neuro-Fuzzy model is very well established approach and has a tremendous potentiality to outcome results with high accuracy ratio and the efficiency with biological data to determine the measure score of matching DNA sequencing. Those recommended sequence-matching algorithm utilize the three input variables – match score (match), mismatch score (mismatch), and gaps, as shown in "Figure 1."

These three inputs would then fuzzified utilizing the following membership functions (MFs) equations and giving the calculated resulting score:

$$\text{Matching} = \begin{cases} 0 & \text{if there is no similarity} \\ 1 & \text{if there is highest similarity 100\%} \\ [1,0] & (\text{matching score} / \text{lenseq}) \end{cases} \quad (1)$$

$$\text{Mismatch} = \begin{cases} 0 & \text{if there is no mismatch} \\ 1 & \text{if there is no similarity} \\ [1,0] & (\text{mismatching score} / \text{lenseq}) \end{cases} \quad (2)$$

$$\text{Gap} = \begin{cases} 0 & \text{if there is no need to put a gap} \\ [1,0] & (\text{gaps score} / \text{lenseq}) \end{cases} \quad (3)$$

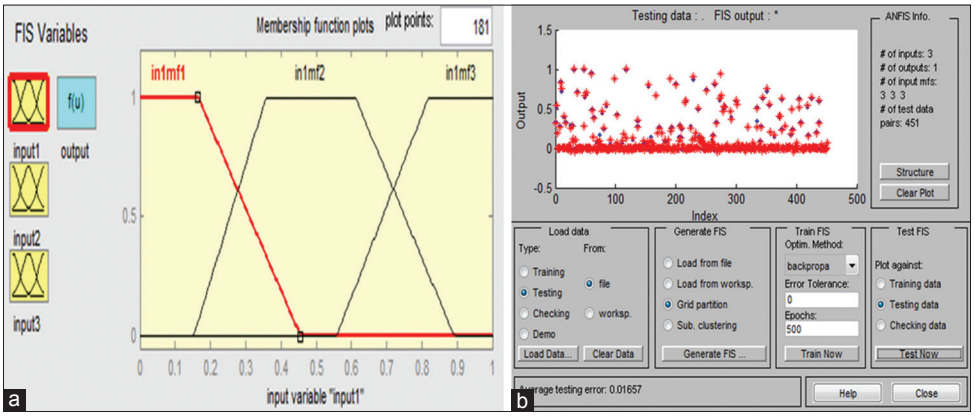

Figure 2: The membership function and the training testing phase for the lowest possible error. (a) Input membership function. (b) The testing data

Table 1: The various ANFIS testing results

Sequences	Global alignment	Identities (%)	Score
AGGTTGC	AGGTTGC	7-May	0.149
AGGTC	AGGT--C	−71%	
GTAGGCTTAAGGTTA	GTAGGCTTAAGGTTA	15-May	0
TAGATC	A- - - T-C- TAG - - - -	−33%	
AGTCCA	A – GTCCA	7-May	0.149
ATGTCC	ATGTCC-	−71%	
CGGGA	CGGGA-	6-Feb	0
ATTGAC	ATTGAC	33%	
CTATCCG	CTATC CG	7-May	0.425
CTAGTCG	CTAGTCG	−71%	

ANFIS: Adaptive neuro-fuzzy inference system

Table 2: The score matching for global alignment of ANFIS tool

Sequences	Local alignment	Identities (%)	Score
AGGTTGC	AGGT	100	1
AGGTC	AGGT		
GTAGGCTTAAGGTTA	TAG	100	1
TAGATC	TAG		
AGTCCA	GTCC	100	1
ATGTCC	GTCC		
CGGGA	GA	100	1
ATTGAC	GA		
CTATCCG	CTA	100	1
CTAGTCG	CTA		

ANFIS: Adaptive neuro-fuzzy inference system

$$\text{Score} = \begin{cases} 0 & \text{if the resulting score} \pounds 0 \\ 1 & \text{if the resulting perfect} \\ [1,0] & (\text{resulting score / perfect score}) \end{cases} \quad (4)$$

The variable "lenseq" mean the entire length of the sequence.

$$Score = match + (-mismatch) + (-gabs) \quad (5)$$

THE SIMULATION RESULTS

In our work, we perform the Neuro-Fuzzy model by an adaptive neuro-fuzzy inference system (ANFIS) tool in Matlab, using the data set about 600 samples for a matching measure score of DNA sequencing; these data divide into two data files for the training and testing, for the training step, we use the data set about 450 samples, and for the testing step, we use the data about 150 samples.[14]

We use these dataset files in ANFIS system in the range value in equation 1, 2, 3, and 4 and output the result according to the equation 5. We use different processing systems to implement the matching process, and each system has different results with convergent values, and that for choosing the most suitable one with less error percentage and depends on it to calculate the matching score. "Figure 2" shows the chosen attempt that gives the results with high accuracy.

"Figure 2" explains the most suitable ANFIS system with the lowest average testing error; Table 1 shows the different processing systems; we implement it with the details. In this table the most suitable system which has chosen is the system that has the following: trapezoidal MFs that has three MFs, constant MFs output, backpropegation train Fuzzy inference system method and the number of epochs which are 500, this system has the lowest average training and testing error which are 0.016572 and 0.01657 respectively, and gives the result with high performance, thus we use it to get the score matching of the numeric data from the DNA sequence alignment, show Tables 2 and 3.

In Table 3, we aligned the sequence using the local algorithm; in this method, the algorithm takes the most similar part of the pair sequence, not must in the order and not need to input the gap; the resulting score is the perfect, there are no mismatch and no gaps, and is 100% identical.

DISCUSSION

Sequence alignment is a necessary condition for analyzing DNA sequencing; in our method, we use the pairwise sequence alignment; it is applied using the global and local alignment algorithm method. In this method, we use the numeric biological data for sequence alignment using ANFIS system in MATLAB; this system implements several processing systems to get the most suitable results as shown in Table 1; the most suitable system is the trapezoidal with three MF for each input and 500 epochs; it has the lowest average testing error. We use several different sequences to be aligned, as shown in Table 2; we aligned the sequences in the global alignment algorithm. In this method, we insert the gaps when the base in the sequence is not similar with the other in the same order in this pair, and shifted the character and input the gap in order to be identical; we use the number of times for (matching), mismatching and gaps as the input in the ANFIS view tool, and output the score matching, as shown in "Figure 3" we can compute the percentage similarity of the alignment by using this code in Matlab [Score, Alignment] = nwalign('S1','S2'); showalignment(Alignment); in order to display a pairwise sequence alignment, as shown in "Figure 4." This use the

Table 3: The score matching for local alignment of ANFIS tool

MF type input	The number of MF for the inputs	The MF output	Train FIS method	The number of epochs	The average training error	The average testing error
Triangular MF	3 3 3	Constant	Backpropagation	50	0.04657	0.06078
Triangular MF	5 5 5	Constant	Backpropagation	250	0.01109	0.04306
Trapezoidal MF	3 3 3	Constant	Backpropagation	100	0.01898	0.0453
Trapezoidal MF	3 3 3	Constant	Backpropagation	500	0.01657	0.01657
Gaussian MF	5 5 5	Constant	Backpropagation	350	0.01436	0.04436
Gaussian2 MF	5 5 5	Constant	Backpropagation	500	0.02804	0.05415

MF: Membership functions, ANFIS: Adaptive neuro-fuzzy inference system, FIS: Fuzzy inference system

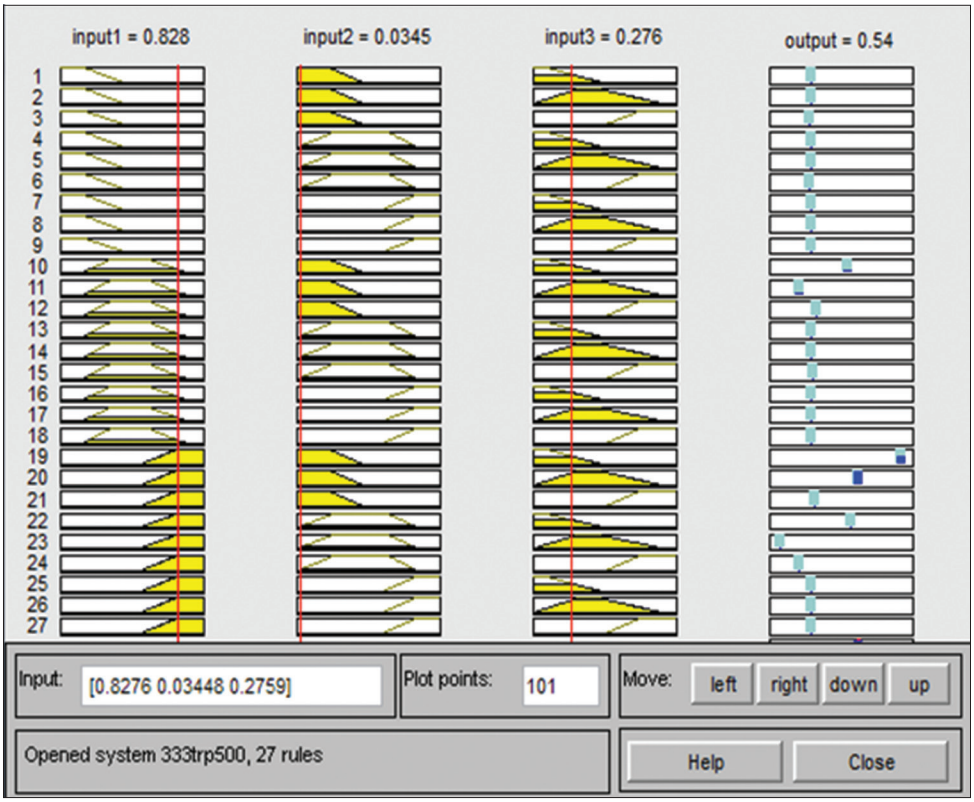

Figure 3: View rules adaptive neuro-fuzzy inference system

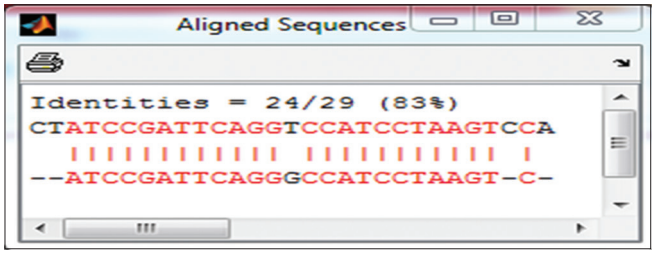

Figure 4: The identical alignment

number of times similar in the sequence alignment divided on the entire length of the alignment sequence.

In Table 3, we aligned the sequence using the local algorithm, in this method the algorithm take the most similar part of the pair sequence, not must in the order and not need to input the gap, the resulting score is the perfect there is no mismatch and no gaps, and have 100% identical.

CONCLUSION

The proposed technique, here, is used to obtain the similarity measure of the pairwise DNA sequence alignment; the pattern matching is an essential task of example the disclosure process in this day and age for finding the basic and utilitarian conduct in the DNA sequencing. In spite of the fact that, the example of matching is commonly utilized as a part of computer science and information processing. In this paper the proposed algorithms are used that are Global and Local alignment to measure the score matching, which are utilized; it as a method to be align the two DNA sequences. The Neuro-Fuzzy model is used to evaluate the score similarity by the ANFIS tool in

Matlab; we implement the method in several processing systems and depend on the most suitable system with the lowest average testing error; we obtain the score matching result for several patterns of DNA sequencing; this model presented the matching implementation in fast and efficient.

REFERENCES

1. R. Bhukya and D. V. L. Somayajulu. "Exact multiple pattern matching algorithm using DNA sequence and pattern pair". *International Journal of Computer Applications*, vol. 17, no. 8, pp. 32-38, 2011.
2. X. Xie, J. Guan and S. Zhou. "Similarity evaluation of DNA sequences based on frequent patterns and entropy". *BMC Genomics*, vol. 16, no. 3, p. S5, 2015.
3. W. Deng and Y. Luan. "Analysis of similarity/dissimilarity of DNA sequences based on chaos game representation". *Abstract and Applied Analysis*, vol. 2013, p. 926519, 2013.
4. P. Pandiselvam, T. Marimuthu and R. Lawrance. "A Comparative Study on String Matching Algorithms of Biological Sequences". In: International Conference on Intelligent Computing, pp. 1-5, 2014.
5. T. Chakrabarti, S. Saha and D. Sinha. "DNA multiple sequence alignment by a hidden markov model and fuzzy levenshtein distance based genetic algorithm". *International Journal of Computer Applications*, vol. 73, no. 16, pp. 26-30, 2013.
6. N. Gill and S. Singh. "Biological sequence matching using fuzzy logic". *International Journal of Scientific and Engineering Research*, vol. 2, no. 7, pp. 1-5, 2011.
7. S. B. Needleman and C. D. Wunsch. "A general method applicable to the search for similarities in the amino acid sequence of two proteins". *Journal of Molecular Biology*, vol. 48, no. 3, pp. 443-453, 1970.
8. T. F. Smith and M. S. Waterman. "Identification of common molecular subsequences". *Journal of Molecular Biology*, vol. 147, no. 1, pp. 195-197, 1981.
9. D. Nauck, F. Klawonn and R. Kruse. "*Foundations of Neuro-Fuzzy Systems*". John Wiley and Sons, Inc., New York, 1997.
10. S. Nasser, G. L. Vert, M. Nicolescu and A. Murray. "Multiple Sequence Alignment Using Fuzzy Logic. In: 2007 IEEE Symposium on Computational Intelligence and Bioinformatics and Computational Biology, IEEE, pp. 304-311, 2007.
11. K. Kim, M. Kim and Y. Woo. "A DNA sequence alignment algorithm using quality information and a fuzzy inference method". *Progress in Natural Science*, vol. 18, no. 5, pp. 595-602, 2008.
12. N. Chai, L. R. Swem, M. Reichelt, H. Chen-Harris, E. Luis, S. Park and J. McBride. "Two escape mechanisms of influenza a virus to a broadly neutralizing stalk-binding antibody". *PLoS Pathogens*, vol. 12, no. 6, p. e1005702, 2016.
13. S. A. Hameed and R. I. Hamed. "Analysing the score matching of dna sequencing using an expert system of neurofuzzy". *Journal of Theoretical and Applied Information Technology*, vol. 95, no. 6, pp. 1255-1262, 2017.
14. DNA Matching Data Base-NCBI". https://www.ncbi.nlm.nih.gov/nucleotide.