A NOVEL APPROACH IN SEARCHING OF QUERY PATTERN IN THE GIVEN TARGET DNA DATABASE USING ARTIFICIAL INTELLIGENCE TECHNIQUE

B. Mukunthan

ABSTRACT

In genetic engineering, the advent of human genome project immensely increased the pressure for molecular computations dealing with data beyond the current abilities that are to be identified and interpreted. The automation of DNA feature extraction process achieved by applying neural network technique which has the advantage over conventional programming, in their ability to solve problem that do not have an algorithmic solution or the available solutions is too complex to be found is discussed in this paper, This work also reduces the complication in precisely analyzing and interpreting human DNA. In this novel approach the perfect blend made of bioinformatics and neural networks results in efficient DNA pattern analysis algorithm with an improved efficiency of 71.25% when compared to an algorithm used in existing tool for pattern searching.

Keywords *–Competitive learning, NFPR-processor, Input Generator, Preprocessor, Separator, Discriminator and Comparator, DNA profiling, DNA sequence Format, Target DNA database, Query database.*

I. INTRODUCTION

Knowledge of DNA sequences has become indispensable for basic biological research. Neural networks learn by examples so that it can be trained with known examples of a problem to gain knowledge about it so the neural network can be effective to solve unknown or untrained instances of the problem if is aptly trained. A pattern [1] [12] is essentially an arrangement or an ordering, in which some organization of underlying structure can be said to exist; a pattern can be referred to as a quantitative or structural [5] description of an object or some item of interest. A set of patterns that share some common properties can be regarded as pattern class [8]; in this work identification numbers generated from nucleotide sequences of the given Human DNA sample. The concept of applying artificial neural systems or artificial neural networks [4] or simply neural networks in the field of DNA profiling is discussed in this paper.As DNA technology is constantly evolving with new techniques, the ability to analyze biological data samples such as DNA or RNA or protein sequences of varying size through classification or recognition requires increased automation, to promise faster and more discriminating results. The need for the proposed system is to design an automated and an efficient algorithm using neural network technique for classification of DNA/RNA obtained from various biological sources.

Assistant Professor, Department of Computer Applications,
Karpagam University, Coimbatore-64102
E-Mail:dr.mukunthan.bmk@gmail.com

II. Methodology

The existing system implemented using conventional algorithms only have the ability to deal with any one type of biological data for what it's designed, at a time either DNA or RNA, but the proposed system can be used to pattern recognize DNA/RNA if it's prior trained with a suitable learning (training) inputs, also there is no need to repeatedly train the system for every sample of similar data set. For instance to pattern recognize various DNA/RNA samples it's sufficient to train the system once with DNA/RNA learning inputs. The Neural-Fuzzy processor implemented in NFPR (Neural-Fuzzy Pattern Recognition) system using the concept of artificial neural network [6] gives novelty to the proposed system with the features such as learning (training) and inference, to classify or recognize patterns which are a perfect reproduction of the training data set comprising of nucleotide base of DNA/RNA which is not present in existing system. The main objective of the proposed system is to effectively perform pattern searching in order to check whether the query pattern is present in the given target DNA database.

III. Literature Review

Neural Networks [3] can process information in parallel, at high speed, and in a distributed manner. Neural networks which are simplified models of the biological neuron system, is a massively parallel distributed processing system made up of highly interconnected neural computing elements that have the ability to learn and thereby acquire knowledge and make it available for use. Neural Network architectures have been classified into various types based on their learning mechanisms and other features. Some classes of Neural Network refer to this learning process as training and the ability to solve a problem using the knowledge acquired as inference.

Neural Networks exhibit mapping capabilities; they can map input patterns to their associated output patterns. Neural Networks architectures [2] [7] can be trained with known examples of a problem before they are tested for their inference. They can, therefore, identify new objects previously untrained. Neural Networks possess the capability to generalize they can predict new outcomes from past trends. Neural Networks are robust systems and are fault tolerant. They can therefore, recall full patterns from incomplete, partial or noisy patterns. In Competitive Learning method [5] [7] the neurons which respond strongly to input stimuli have their weights updated, when an input pattern is presented, all neurons in the layer compete and the winning neuron undergoes weight adjustment. Hence it is a "Winner-takes-all" strategy.

DNA profiling [17] also called DNA testing, DNA typing [13] [15], or genetic fingerprinting, is a technique employed by forensic scientists [14] [11] to assist in the identification of individuals on the basis of their respective DNA profiles. DNA profiles [17 [18], are encrypted sets of numbers that reflect a person's DNA makeup, which can also be used as the person's identifier. DNA sequencing theory addresses physical processes related to sequencing DNA[9] [10] .The term DNA sequencing [19] refers to sequencing methods for determining the order of the nucleotide bases—adenine, guanine, cytosine, thymine and uracil (rare case) in a molecule of DNA. Single nucleotide poly-orphisms [20] are a DNA sequence variation occurring when a single nucleotide A, T, C, or G in the genome (or other shared sequence) differs between members of a species [16] (or between paired chromosomes in an individual). The genome [21] is the entirety of an organism's hereditary information which is encoded either in DNA or, for many types of virus, in RNA. For instance, two sequenced DNA fragments from different individuals, AAGCCTA to AAGCTTA, contain a difference in a single nucleotide.

Various representations of nucleotides present in DNA and RNA is shown in Table 1.

Table 1 : Representation of nucleotides present in DNA and RNA

S.No	Nucleotides	Presence	Character & Color Representation		Fuzzy Equivalent
1	Adenine	DNA/RNA	A	■	0.1
2	Thymine	DNA	T	■	0.2
3	Guanine	DNA/RNA	G	■	0.3
4	Cytosine	DNA/RNA	C	■	0.4
5	Uracil	RNA	U	■	0.5

IV. Pattern Searching Using Hash Coding Technique

The hash coding algorithm is used in very widely used sequence matching and search programs BLAST and FASTA available in modern tools .In hash coding method if the k-tuple is large, the speed is high, specificity, that is the ability to pick up accurate and meaningful matches is high and the sensitivity, that is the approximate or distant matching is low. Conversely, if the k-tuple is small, the speed is low, the specificity is low, but the sensitivity is high. In these case since the comparison is executed in each repetition of the loop, the comparison operation is considered as the algorithm's base operation.

A. Analysis of Hash coding algorithm in pattern searching

Target Database:-

Let c1 and c2 are the count of base operation performed in the hash table for k-tuple (k=1) that is A, T, G, and C and for k-tuple (k=2) that is AA, AT, AG, AC… CC respectively. For target input of size 43

$$c1\ (43) = 172 \text{ and } c2\ (43) = 172$$

Count of base operation for target database: $ct\ (43) = c1\ (43) + c2\ (43) = 344$

Query Pattern:-

Let c3 and c4 are the count of basic operation performed in the hash table for k-tuple (k=1) that is A, T, G, and C and for k-tuple (k=2) that is AA, AT, AG, AC… CC respectively. For query pattern of size 7

$$c3\ (7) = 28$$

$$c4\ (7) = 28$$

Count of base operation for query pattern

$$cq\ (7) = c3\ (7) + c4\ (7) = 56$$

The total count of base operation in hash coding algorithm for input of size 50, $C_{hash}(50)$ is given by

$$C_{hash}(50) = ct\ (43) + cq\ (7)$$

$$C_{hash}(50) = 344 + 56$$

$$C_{hash}(50) = 400$$

The run time efficiency of hash coding algorithm, for the input of size 50, $T_{hash}(50)$ is given by

$$T_{hash}(50) \approx C_{hash}(50)=400/100 \approx 4.00 \text{ Seconds}$$

Where, one base operation = 1/100 th of a second.

V. Proposed Neural-Fuzzy Pattern Recognition System [NFPR]

Pattern classification is done in the given Human DNA/ RNA target database in order to check whether the query pattern is found in it. The automation of generating identification numbers from a given target DNA sample can be achieved using the proposed system by classifying the sequence into valid sequences and invalid sequences. The identification numbers of valid sequences are compared with the identification number of the query pattern through which the presence or absence of query pattern in target DNA database is confirmed.

Various components of NFPR system and its functions are shown in figure 1

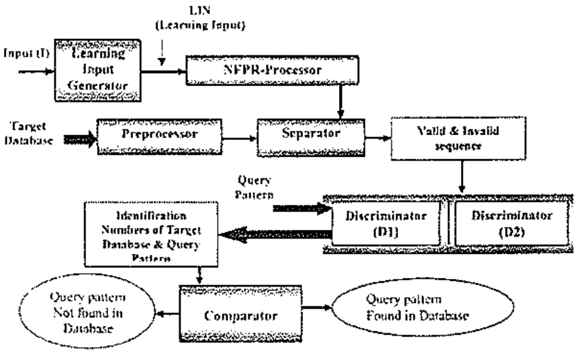

Figure 1 : Neural-Fuzzy Pattern Recognition System (NFPR)

Learning Input Generator

The input normalization process is essential since Neural-Fuzzy pattern recognition system needs all the input values to lie between 0 and 1.

$$\text{Learning Inputs [NFPR]: } LIN_{i,n} = I1, I2..., Ip$$

$$\text{Where } 0.1 \leq i \leq 0.5, \quad 0.1 \leq n \leq 0.5 \text{ and } p= 4$$

Neural-Fuzzy [NF] Processor

The normalized learning inputs from learning input generator is fed to the Neural-Fuzzy processor to compute two major functions such as ignition function (IGF) and tracking function (TRF) that assist to generate weights for inference(WFI), category for inference(CFI) of the inputs that are trained.The tracking function also helps to determine whether the network must adjust its learning parameters.

Preprocessor

The preprocessor of NFPR system receives a nucleotide pair input of the DNA sample for which the identification numbers is to be generated and a nucleotide pair of target database in which the query pattern is to be searched, to produce an output in the format suitable for separator based on the various conditions .

Separator

The separator of NFPR computes a category inference function (CIF) of the entire preprocessed sample from preprocessor for all weights for inference (WFI) and category for inference (CFI) obtained from the Neural-Fuzzy processor.

The separator of NFPR also identifies greatest inferred category (GIC), from which the category it belongs to is determined, if the category is valid (v) five consecutive nucleotide base after the nucleotide pair in the sample is considered as a sequence of valid category and if the category is invalid the two nucleotide base in the pair is considered as a sequence of invalid category.

Discriminator [D1, D2]

The separator outputs which are valid in their category are fed to the discriminator (D1) where the corresponding identification numbers are computed. The separator outputs which are invalid in their category are fed to the discriminator (D2) where the identification numbers of the given sample is computed. The output (D1) is used for pattern searching.

Comparator

The comparator unit of NFPR compares the identification number of valid sequences of target data base with identification number of query pattern and confirms whether query pattern is available in the given database.

A. Pattern Searching in the DNA sequence using NFPR technique

Step1: Learning inputs to the Neural-Fuzzy processor is normalized using various condtions as given in Table 2.

Step2: Generating Weight for Inference and Categoy for Inference using Ignition Function (IGF) and Tracking Function (TRF).

Step3: Inference of Category for the nucleotide pairs in 'Target database' shown in figure 2 is done using Category Inference Function (CIF).

Table 2 : Conditions for learning input normalization in NFPR system

	Condition	Learning Input	Category
Case I	i ≠n (or) i=n=0.1 and n<=0.5	$LIN_{i,n}$ = i, n, 1- i, 1-n e.g. $LIN_{0.1,0.1}$ = 0.1, 0.1, (1- 0.1), (1-0.1) $LIN_{0.1,0.1}$ = 0.1, 0.1, 0.9, 0.9 $LIN_{0.2,0.5}$ = 0.2, 0.5, (1- 0.2), (1-0.5) $LIN_{0.2,0.5}$ = 0.2, 0.5, 0.8, 0.5	v(valid)
		....	
Case n	i =n=0.5	$LIN_{i,n}$ = i, i+0.1, n, n-0.1 e.g. $LIN_{0.5,0.5}$= 0.5, (0.5+0.1), 0.5, (0.5-0.1) $LIN_{0.5,0.5}$= 0.5, 0.6, 0.5, 0.4	i (invalid)

Step4: Generate unique identification number for all valid sequence in 'Target DNA Database' [Base pair=32,Sequence=25] as shown in figure 3, using

equation , $$D1_{t,s} = \sum_{k=1}^{7} k(Vseq_{t,s,k})^{k}$$

Where,Vseq=valid sequence, t=target database, s=sequence, k=position of nucleotide base in sequence.

Step5: Generate unique identification number for the 'Query Pattern' as shown in

figure 3 using the above same equation.

Step6: Compare the identification number of Query Pattern with all the identification number of Target database and if there is a match then the pattern is confirmed in the database, if no match the absence of pattern in the database is confirmed.

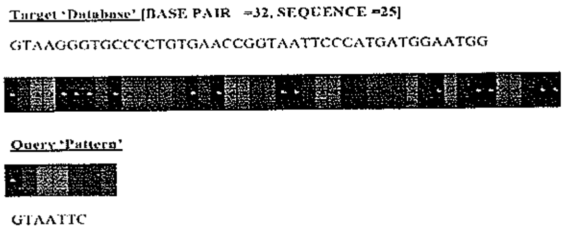
Target 'Database' [BASE PAIR =32, SEQUENCE =25]
GTAAGGGTGCCCCTGTGAACCGGTAATTCCCATGATGGAATGG
Query 'Pattern'
GTAATTC

Figure 2 : Sample of pattern searching

The Separator Output of NFPR Processor for Target Database is shown in Figure 3.

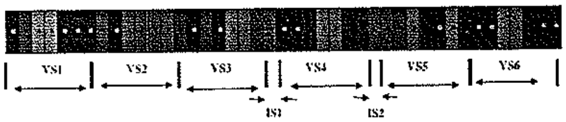

Figure 3 : Separator output of NFPR processor

Discriminator outputs of '*Target database*':

$D1_{1,1}$ of VS1 =0.4014549

$D1_{1,2}$ of VS2 =0.7502656

$D1_{1,3}$ of VS3 =0.4974948

$D1_{1,4}$ of VS4 =**0.4004528**

$D1_{1,5}$ of VS5 =0.4783669

$D1_{1,6}$ of VS6 =0.3459240

Discriminator output of '*Query Pattern*':

$D1_{query}$=**0.4004528.**

B. Analysis of Proposed algorithm in pattern searching

Let c1 is the count of base operation performed in the learning (training) and inference, c2 is the count of base operation performed in comparison of identification numbers in proposed NFPR algorithm.

For target input of size 50, c1 (50) = 27+82=109 and c2 (50) = 6

The total count of base operation in proposed algorithm for input of size 50, $C_{nfpr}(50)$ is given by $C_{nfpr}(50)$ = c1 (50) + c2 (50);$C_{nfpr}(50)$ =109+6 =115

The run time efficiency of proposed algorithm, for input of size, $T_{nfpr}(50)$ is given by $T_{nfpr}(50) \approx C_{nfpr}(50)$ =115/ 100 $\approx$1.15 Seconds

Where, one base operation = 1/100 th of a second.

VI. Performance Evaluation of Existing Algorithm versus Proposed Algorithm in Pattern Searching

The learning time required for number of epochs of proposed algorithm during learning phase is given in table 3.

Table 3 : Performance of proposed system for various numbers of epochs

S. No	Learning Vector (Number of Epochs)	Number of Learning Inputs	Learning Time (Seconds)
1	25	25	62.49
2	13	13	34

No. of Epochs=25(For All Possible Combinations) & No. of Epochs =13(Proposed Algorithm)

The efficiency of the existing hash coding algorithm and proposed NFPR algorithm is measured using the base operation performed in the algorithm. The existing system requires 400 comparisons with time frame of 4.00 seconds (one base operation=1/100 seconds), but the proposed system only requires 1.15 seconds for 115 comparisons, as shown in Mat lab output of Figure 4 and Figure 5.

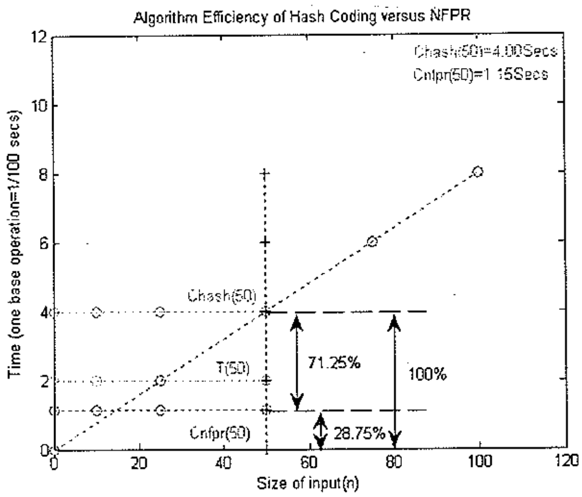

Figure 4 : Base operation of Hash coding algorithm and NFPR for input size n=50

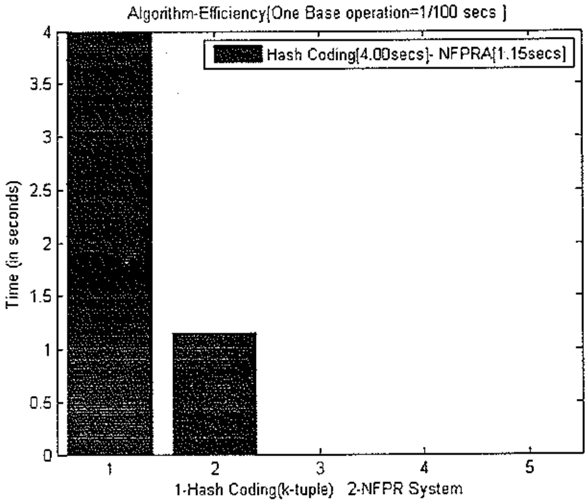

Figure 5 : Algorithm efficiency of NFPR versus Hash coding

VII. Conclusion

The proposed algorithm implemented using the neural network technique classifies the human DNA sequences to generate identification numbers which assists to efficiently search query pattern in the given DNA target database with an improved efficiency of 71.25% over conventional algorithm implemented in the searching tool. Further development can be done in the above work to implement it in protein folding problem by training patterns sequences using suitable fuzzy equivalent and also above technique can be used in the areas where feature extraction and classification is to be done in genetic engineering with suitable modification.

References

1. Richard O. Duda, Peter E.Hart, David G. Stork, *"Pattern classification"-Second Edition"*, John Wiley and sons, 2012.

2. John Hertz, Anders Krogh, and Richard G. Palmer. *"Introduction to the Theory of Neural Computation"*. Addison Wesley, Redwood City, A, 2008.

3. Stephen J. Hanson, Jack D. Cowan, and C. Lee Giles, *"Advances in Neural Information Processing Systems"*, volume 5, Morgan Kaufmann San Mateo CA, 2009.

4. *"Advances in Neural Networks issn-2006"*, Third international symposium on neural networks, Springer Berlin Heidelberg, New York publications.

5. Robert Schalkoff, *"Pattern Recognition: Statistical, Structural and Neural Approaches, 2012,* John Wiley and sons.

6. Carpenter, G.A. and S. Grossberg, *"A Massively Parallel Architecture for a self-organizing Neural Pattern Recognition Machine"*, Computer Vision, Graphics and Image Processing, 37, PP. 54-115.

7. Carpenter, G.A. and S. Grossberg, and J.H. Reynolds (2012), *"ARTMAP: Supervised Real Time Learning and Classification of Non-stationary Data by a Self- organizing Neural Network"*. Vol. 4, pp. 565-588.

8. Phipps Arabie, Lawrence J. Hubert, and Geert De Soete, editors, *"Clustering and Classification"*. World Scientific, River Edge, NJ.

9. Stephen,Krawetz, David D.Womble , *"Introduction to Bioinformatics A Theoretical and Practical Approach"*, Human Press Inc,2003.

10. David W.Mount, David W. Mount, *"Bio informatics Sequence and Genome analysis"*- Second Edition, Cold Spring Harbor Laboratory Press, New York,2005.

11. Norah Rudin, Keith Inman, *"An Introduction forensic DNA Analysis"*, CRC Press, 2011.

12. Donald R. Tveter. *"The Pattern Recognition Basis of Artificial Intelligence"*. IEEE Press, New York, page 117, Computational Intelligence and Bio inspired Systems, 8th international work conference on artificial neural networks, iwann-2005proceedings.

13. Julie A. Ayala-Gross, *"DNA Analysis: The best method for Human Identifications"*, National University, San Diego – 2001.

14. Joe Nickell and John F.Fischar, *"Crime Science Methods of Forensic Detection"*, 1999. University Press of Kentucky.

15. David E. Newton, *"DNA Evidence and Forensic science"*- 2008 facts on file, Inc. http:// www.factsonfile.com.

16. Jorg T. Epplen Thomas Lubjuhn, Birkhauser, *"DNA Profiling and DNA Finger Printing"*, Verlag Publication,1999.

17. Simon Eastaeal, Neil Mc Lead, Ken, Harwood , *"DNA Profiling Principles, Pitfalls and Potential"*, Academic Publishers, Inc,1991.

18. Des Higgins, willie Taylor, *"Bioinformatics Sequence, Structure and data banks"*, Oxford University Press, 2000.

19. *"Bioinformatics for geneticists"*, Michael R.Barnes , Second Edition, John Wiley & Sons Ltd.,

20. Andreas D. Buxevanis, *"Bioinformatics-A practical Guide to the Analysis of genes and proteins"*, second edition, A John wiley & sons, Inc., Publication, 2002.

21. S.N Sivanandam, *"Introduction to neural networks and MATLAB-6.0"*, Tata McGraw-Hill publishing company, 2006.

AUTHOR BIOGRAPHY

Dr. B. Mukunthan., was born on 08th of May 1977 in Coimbatore, Tamil Nadu, India. He completed his post graduation in computer applications at Coimbatore Institute of Management and Technology, Coimbatore and his PhD in Computer Applications from Anna University-Chennai. He is also a Microsoft Certified Professional, Microsoft Certified Application Developer (MCAD) and Microsoft Certified Solution Developer (MCSD) in .NET Technology. He has five international journal publications and presented four national and international conference papers in his Research area. His area of interest is artificial intelligence, neural networks and Fuzzy logic and Data Mining.