Automation of DNA Finger Printing for Precise Pattern Identification using Neural-fuzzy Mapping approach

A. Pushpalatha
Prof., Dept of Mathematics,
Govt Arts College, Udumalpet,
Coimbatore-642126,
Tamil Nadu, India

B. Mukunthan
HOD, Dept. of Master of Computer Applications,
SVS College of Engineering,
Coimbatore -642109,
Tamil Nadu, India

ABSTRACT

The conventional techniques and algorithms employed by forensic scientists to assist in the identification of individuals on the basis of their respective Deoxyribonucleic acid base(DNA) pair profiles involves more computational steps and mathematical formulas that leads to more complexity. DNA identification is not considered by many as a biometric recognition technology, mainly because it is not yet an automated process i.e. it takes more time to analyze the DNA finger prints and samples collected from the crime scene, it will be considered as a future biometric trait if it's suitably automated. 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 it is aptly trained. The perfect blend made of bioinformatics, neural networks and fuzzy logic results in efficient algorithms of pattern analysis techniques that induce automation which is inevitable in DNA profiling that became manually impractical with the growing amount of data.

Keywords

Competitive Learning, DNA Profiling, Adaptive Resonance Theory, Simplified Fuzzy ARTMAP, Input Generator, DNA Sequence Format, Input Preprocessor, Input Separator, and Deviator.

1. INTRODUCTION

Knowledge of DNA sequences has become indispensable for basic biological research. DNA profiling is applied in various fields such as diagnostic, biotechnology, forensic biology and biological systematic.

The DNA sequences of thousands of organisms have been decoded and stored in databases. The sequence information is analyzed to determine genes that encode polypeptides, RNA genes, regulatory sequences, structural motifs, and repetitive sequences. A comparison of genes within a species or between different species can show similarities between protein functions, or relations between species.

The rapid speed of sequencing attained with modern DNA sequencing technology has been instrumental in the sequencing which assists in DNA profiling or genetic finger printing of the human genome.

The practical aspects revolve around designing and optimizing DNA profiling projects, predicting project performance, troubleshooting experimental results, characterizing factors such as sequence bias and the generation of software processing algorithms for DNA profiling by analysing sequenced data. The concept of applying Artificial Neural Systems (ANS) or Neural networks in the field of DNA profiling is discussed in this paper.

2. NEURAL NETWORK TECHNIQUES

2.1 Neural Networks

Neural Networks [1] [2] can process information in parallel, at high speed, and in a distributed manner. Learning methods in neural networks are classified as supervised learning, unsupervised learning; reinforce learning, hebbian learning, gradient descent learning, competitive learning and stochastic learning.

In competitive learning method those 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.

Neural networks [3], 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, i.e., they can map input patterns to their associated output patterns so that it can 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.

Neural networks found wide applications in areas such as pattern recognition [17], image processing, optimization, fore casting and control systems.

2.2 Adaptive Resonance Theory

Adaptive resonance theory [5] employs a new principle of self organization based on competitive learning. Adaptive resonance theory nets are designed to be both stable and plastic.

The Neural networks suitable particularly for pattern classification problems in realistic environment is simplified fuzzy ARTMAP [1] [4] [5], it is a vast simplification of fuzzy ARTMAP which has reduced computational overhead and architectural redundancy when compared to its predecessor.

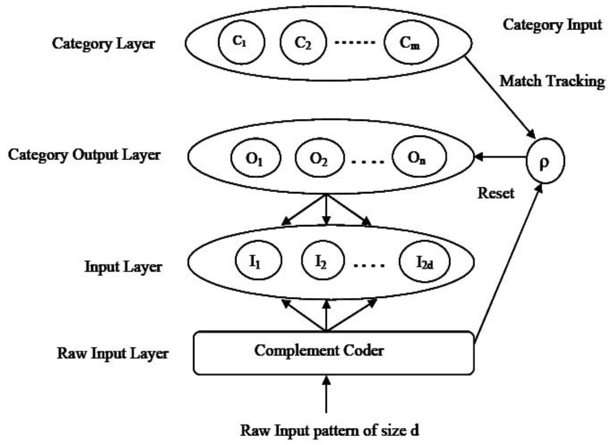


Figure 1 Simplified fuzzy ARTMAP

In the simplified ARTMAP the vigilance parameter (ρ) can range from 0 to 1 that controls the granularity of the output node encoding i.e., High vigilance values makes the output node much fussier during pattern encoding, low vigilance renders the output node to be liberal during the encoding of patterns.

3. DNA PROFILING AND SEQUENCING

3.1 DNA Profiling

DNA profiling[6] [7] also called DNA testing, DNA typing, or genetic fingerprinting is a technique employed by forensic scientists to assist in the identification of individuals on the basis of their respective DNA profiles. DNA profiles are encrypted sets of numbers that reflect a person's DNA makeup, which can also be used as the person's identifier.

DNA profiling is widely used in parental testing and rape investigation. Forensic science [13] [14] often shortened to forensics is the application of a broad spectrum of sciences to answer questions of interest to a legal system that may be in relation to a crime or a civil action. Human DNA [9] [10] sequences are the same in every person; enough of the DNA is different to distinguish one individual from another.

DNA profiling uses repetitive "repeat" sequences that are highly variable, called variable number tandem repeats (VNTR). VNTRs loci are very similar between closely related humans, but so variable that unrelated individuals are extremely unlikely to have the same VNTRs.

3.2 DNA Sequencing

DNA sequencing theory addresses physical processes related to sequencing DNA e.g. sequence alignment. The term DNA sequencing [11] [15] refers to sequencing methods for determining the order of the nucleotide bases—adenine, guanine, cytosine, thymine and uracil in a molecule of DNA.

Single nucleotide poly-orphisms [12] is 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 (or between paired chromosomes in an individual).

The genome [16] is the entirety of an organism's hereditary information which is encoded either in DNA or, for many types of virus, in RNA. For example, two sequenced DNA fragments from different individuals, AAGCCTA to AAGCTTA, contain a difference in a single nucleotide.

The method of DNA sequencing [8][16] separates the DNA sample into four groups each one treated with a specific restriction enzyme for A, T, G, or C. After this, all four groups are placed in the same apparatus for gel electrophoresis resulting in a DNA sequence.

Various DNA Sequence Formats available are 1) Plain sequence format 2) EMBL format 3) GCG format 4) GCG-RSF (rich sequence format 5) GenBank format 6) IG format 7) FASTA format.

FASTA format -A sequence file in FASTA format can contain several sequences each sequence in FASTA format begins with a single-line description, followed by lines of sequence data. It is a text-based format for representing either nucleotide sequences or peptide sequences [6], in which base pairs or amino acids are represented using single-letter codes.

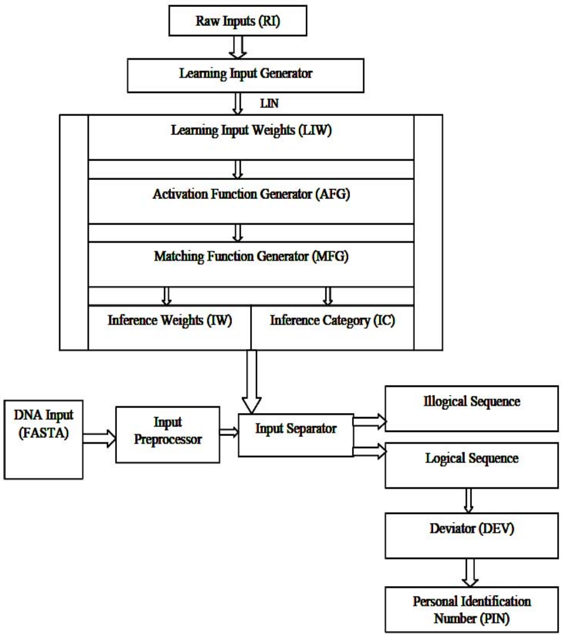


Figure 2 Block diagram of Neural-Fuzzy Pattern Recognition System (NFPRS)

The format also allows for sequence names and comments to precede the sequences. The FASTA format may be used to represent either single sequences or many sequences in a single file.

A series of single sequences, concatenated, constitute a multi sequence file. It is common to end the sequence with an "*" (asterisk) character and leave a blank line between the description and the sequence.

3.2.1 Example sequence in FASTA format

>AB000263 |acc=AB000263|descr=Homo sapiens mRNA for prepro cortistatin like peptide, complete cds.|len=368

ACAAGATGCCATTGTCCCCCGGCCCCTGCTGCTGCTGC
TCTCCGGGGCACGGCCACCGCTGCCCTGCCCCTGGA

GGGTGGCCCCACCGGCCGAGACAGCGAGCATATGCAG
GAAGCGGCAGGAATAAGGAAAAGCAGCCTCCTGAC

TTCCTCGCTTGGTGGTTTGAGTGGACCTCCCAGGCCAG TGCCGGGCCCCTCATAGGAG

4. NEURAL- FUZZY PATTERN RECOGNITION PROCESSOR

4.1 Learning Input (LIN) Generator

The input generator is used for Raw Input (*RI*) normalization and it represents the presence of particular feature in the input patterns and its absence.

$$LIN_{i,\,n} = RI_1, RI_2 \ldots\ldots, RI_p \qquad \text{(Eq.1)}$$

where $0.1 \le i \le 0.5,\ 0.1 \le n \le 0.5$

and $p = 4$

Case 1: $i \neq n$ or $i=n=0.1$ Then LIN $_{i,\ n}$ = i, n, 1- i, 1-n and Category=L (Logical)
e.g., LIN $_{0.1,\,0.1}$ = 0.1, 0.1, (1- 0.1), (1-0.1)
i.e., LIN $_{0.1,\,0.1}$ = 0.1, 0.1, 0.9, 0.9

Case 2: i = n and i, n > 0.1 Then LIN $_{i,\,n}$ = i, 1-i, 1-n, n and Category=ILL (Illogical)
e.g., LIN $_{0.2,\,0.2}$= 0.2, (1-0.2), (1- 0.2), 0.2
i.e., LIN $_{0.2,\,0.2}$ = 0.2, 0.8, 0.8, 0.2

4.2 Activation function (ACF) generator

When coded input patterns from input generator are presented to NFPRS-processor all output nodes become active to varying degrees. The output activation denoted by ACFj is referred to as the Activation Function for the jth output node. Where LIN is the Learning input and LIW$_j$ is the corresponding learning input weights.

$$ACFj = \frac{|LIN \wedge LIWj|}{\alpha + |LIWj|} \qquad \text{(Eq.2)}$$

Here α is kept as a small value close to 0 it's about 0.0000001.The node which registers the highest activation function is deemed winner node i.e.

$$\text{Winner node} = \max(ACFj) \qquad \text{(Eq.3)}$$

In the event of more than one node emerging as the winner owing to the same activation function value, a mechanism such as choosing a node with the smallest index is devised to break the tie. The category associated with the winner is the one to which the given input pattern parameters.

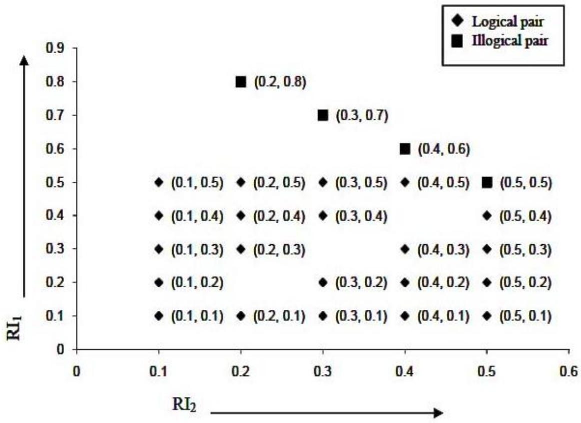
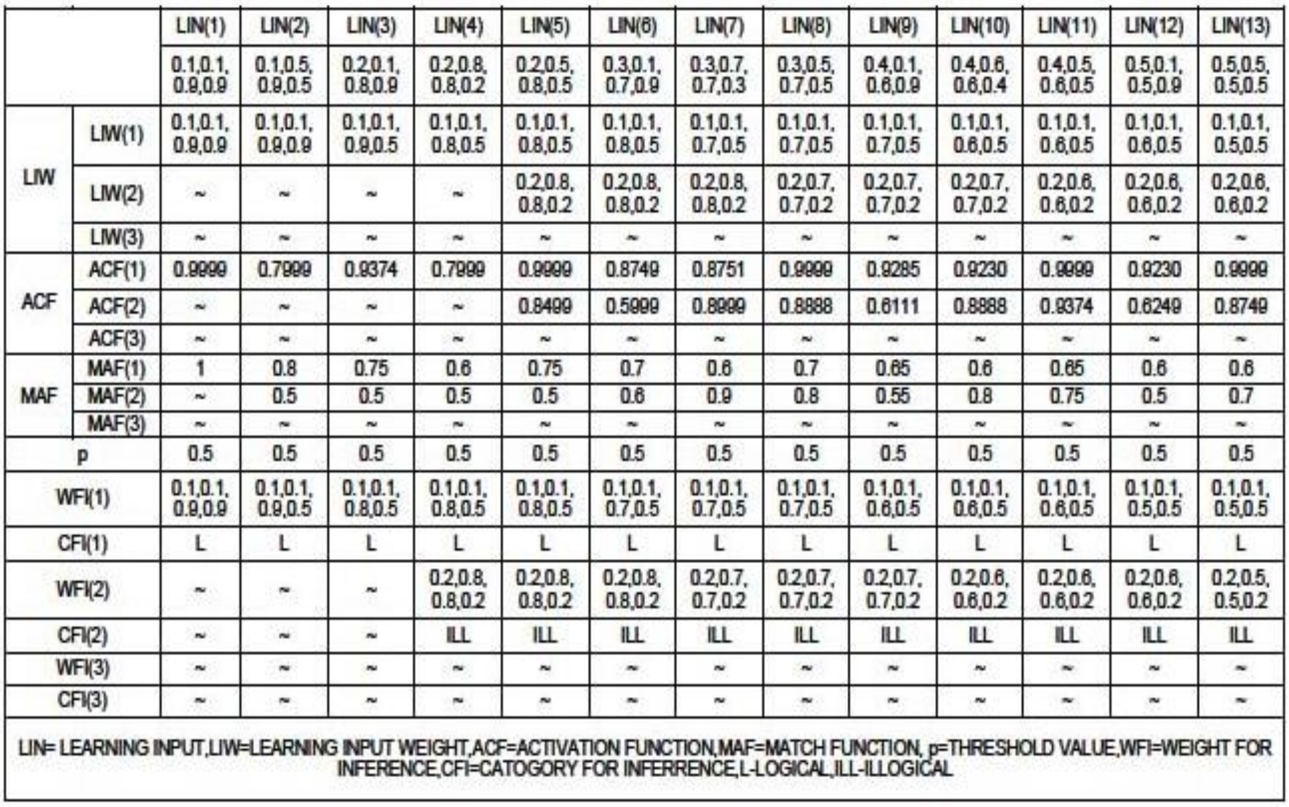


Figure 3 Concept of clustering logical sequence and illogical sequence

Table 1 Generating weights for inference, category for inference from learning inputs



		LIN(1)	LIN(2)	LIN(3)	LIN(4)	LIN(5)	LIN(6)	LIN(7)	LIN(8)	LIN(9)	LIN(10)	LIN(11)	LIN(12)	LIN(13)
		0.1,0.1, 0.9,0.9	0.1,0.5, 0.9,0.5	0.2,0.1, 0.8,0.9	0.2,0.8, 0.8,0.2	0.2,0.5, 0.8,0.5	0.3,0.1, 0.7,0.9	0.3,0.7, 0.7,0.3	0.3,0.5, 0.7,0.5	0.4,0.1, 0.6,0.9	0.4,0.6, 0.6,0.4	0.4,0.5, 0.6,0.5	0.5,0.1, 0.5,0.9	0.5,0.5, 0.5,0.5
LIW	LIW(1)	0.1,0.1, 0.9,0.9	0.1,0.1, 0.9,0.9	0.1,0.1, 0.9,0.5	0.1,0.1, 0.8,0.5	0.1,0.1, 0.8,0.5	0.1,0.1, 0.8,0.5	0.1,0.1, 0.7,0.5	0.1,0.1, 0.7,0.5	0.1,0.1, 0.7,0.5	0.1,0.1, 0.6,0.5	0.1,0.1, 0.6,0.5	0.1,0.1, 0.6,0.5	0.1,0.1, 0.5,0.5
	LIW(2)	~	~	~	~	0.2,0.8, 0.8,0.2	0.2,0.8, 0.8,0.2	0.2,0.8, 0.8,0.2	0.2,0.7, 0.7,0.2	0.2,0.7, 0.7,0.2	0.2,0.7, 0.7,0.2	0.2,0.6, 0.6,0.2	0.2,0.6, 0.6,0.2	0.2,0.6, 0.6,0.2
	LIW(3)	~	~	~	~	~	~	~	~	~	~	~	~	~
ACF	ACF(1)	0.9999	0.7999	0.9374	0.7999	0.9999	0.8749	0.8751	0.9999	0.9285	0.9230	0.9999	0.9230	0.9999
	ACF(2)	~	~	~	~	0.8499	0.5999	0.8999	0.8888	0.6111	0.8888	0.9374	0.6249	0.8749
	ACF(3)	~	~	~	~	~	~	~	~	~	~	~	~	~
MAF	MAF(1)	1	0.8	0.75	0.6	0.75	0.7	0.6	0.7	0.65	0.6	0.65	0.6	0.6
	MAF(2)	~	0.5	0.5	0.5	0.5	0.6	0.9	0.8	0.55	0.8	0.75	0.5	0.7
	MAF(3)	~	~	~	~	~	~	~	~	~	~	~	~	~
p		0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
WFI(1)		0.1,0.1, 0.9,0.9	0.1,0.1, 0.9,0.5	0.1,0.1, 0.8,0.5	0.1,0.1, 0.8,0.5	0.1,0.1, 0.8,0.5	0.1,0.1, 0.7,0.5	0.1,0.1, 0.7,0.5	0.1,0.1, 0.7,0.5	0.1,0.1, 0.6,0.5	0.1,0.1, 0.6,0.5	0.1,0.1, 0.6,0.5	0.1,0.1, 0.5,0.5	0.1,0.1, 0.5,0.5
CFI(1)		L	L	L	L	L	L	L	L	L	L	L	L	L
WFI(2)		~	~	~	0.2,0.8, 0.8,0.2	0.2,0.8, 0.8,0.2	0.2,0.8, 0.8,0.2	0.2,0.7, 0.7,0.2	0.2,0.7, 0.7,0.2	0.2,0.7, 0.7,0.2	0.2,0.6, 0.6,0.2	0.2,0.6, 0.6,0.2	0.2,0.6, 0.6,0.2	0.2,0.5, 0.5,0.2
CFI(2)		~	~	~	ILL	ILL	ILL	ILL	ILL	ILL	ILL	ILL	ILL	ILL
WFI(3)		~	~	~	~	~	~	~	~	~	~	~	~	~
CFI(3)		~	~	~	~	~	~	~	~	~	~	~	~	~

LIN= LEARNING INPUT,LIW=LEARNING INPUT WEIGHT,ACF=ACTIVATION FUNCTION,MAF=MATCH FUNCTION, p=THRESHOLD VALUE,WFI=WEIGHT FOR INFERENCE,CFI=CATOGORY FOR INFERRENCE,L-LOGICAL,ILL-ILLOGICAL

4.3 Match Function (MAF) Generator

$$\text{MAFj} = \frac{\left| \text{LIN} \wedge \text{LIWj} \right|}{\left| \text{LIN} \right|} \qquad \text{(Eq.4)}$$

The match function in association with the monitoring parameter decides on whether a particular output node is to encode a given input pattern or whether a new output node should be opened to encode the same.

The network is said to be in a state of resonance, it is essential that it not only encodes the given input pattern but should also represent the same category as that of the input patterns.

The network is said to be in state of mismatch reset if the monitoring parameter exceeds match function. That means the particular output node is not fit enough to learn the given input pattern and thereby cannot update its weights even though the category of the output node may be the same as that of the input pattern. This is so, since the output node has fallen short of the expected encoding granularity indicated by the monitoring parameter.

The weight updating equation of an output node *j* when it proceeds to learn the given input pattern is given by Weight for Inference (WFI)

$$\text{WFI j}^{\text{new}} = \chi\,(\text{LIN} \wedge \text{WFIj}^{\text{old}}) + (1-\chi)\text{WFIj}^{\text{old}}$$

$$\text{where} \quad 0 \le \chi \le 1 \qquad \text{(Eq.5)}$$

The category activation function (CIF) is given by

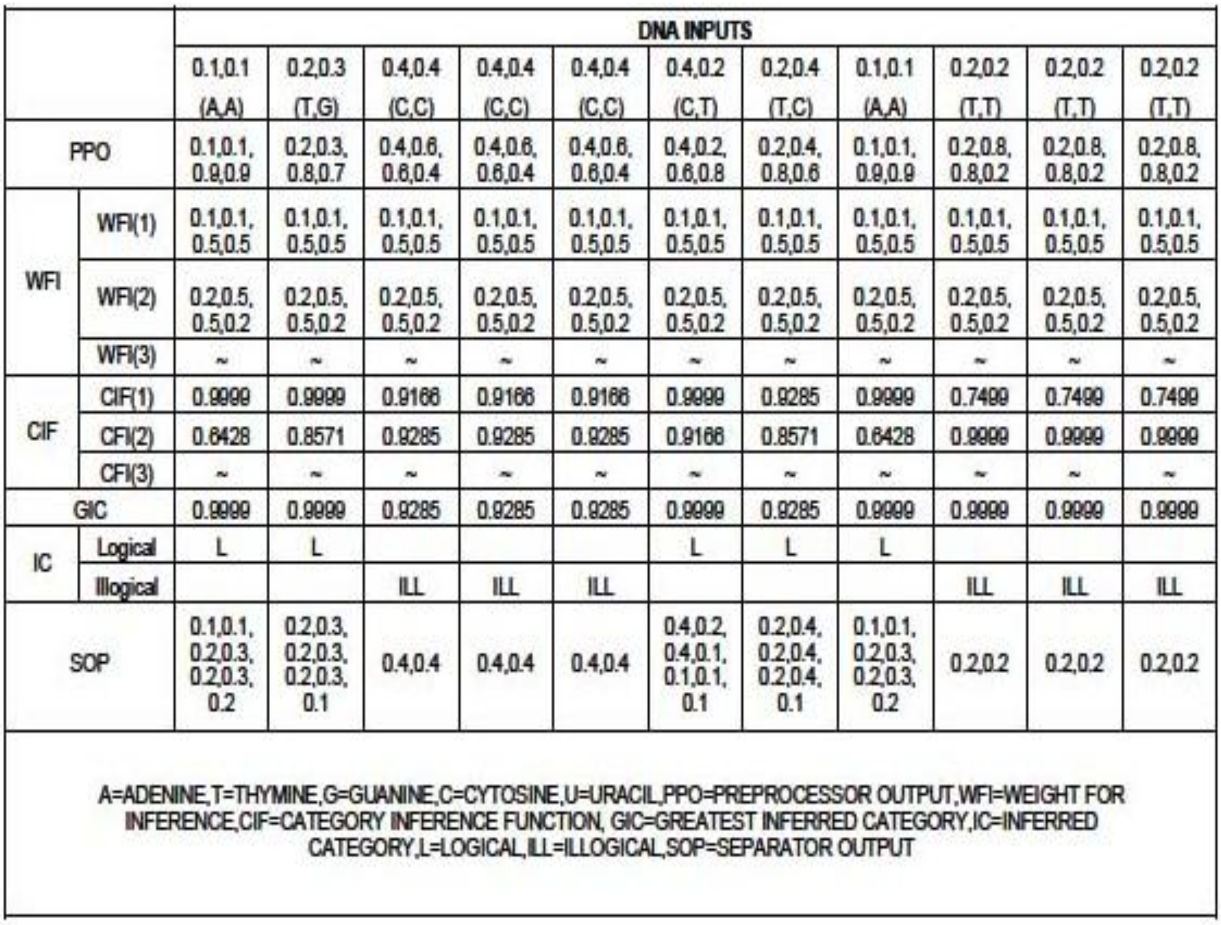
$$\text{CIFj} = \frac{\left| \text{PPO} \wedge \text{WFIj} \right|}{\left| \text{WFIj} \right|} \qquad \text{(Eq.6)}$$

All the output nodes compute the activation functions with respect to the input. The winner, node with the highest activation function, is chosen.

The category to which output node belongs is the one to which given input pattern is classified by the network, where input preprocessor output is denoted by (PPO).

Once the network has been trained the categories to which the patterns belong may be easily computed by inferring weight for inference (WFI) and category activation function (CIF).

Table 2 Inferring categories of Human-1 using weights for inference, category activation function



		DNA INPUTS										
		0.1,0.1 (A,A)	0.2,0.3 (T,G)	0.4,0.4 (C,C)	0.4,0.4 (C,C)	0.4,0.4 (C,C)	0.4,0.2 (C,T)	0.2,0.4 (T,C)	0.1,0.1 (A,A)	0.2,0.2 (T,T)	0.2,0.2 (T,T)	0.2,0.2 (T,T)
PPO		0.1,0.1, 0.9,0.9	0.2,0.3, 0.8,0.7	0.4,0.6, 0.6,0.4	0.4,0.6, 0.6,0.4	0.4,0.6, 0.6,0.4	0.4,0.2, 0.6,0.8	0.2,0.4, 0.8,0.6	0.1,0.1, 0.9,0.9	0.2,0.8, 0.8,0.2	0.2,0.8, 0.8,0.2	0.2,0.8, 0.8,0.2
WFI	WFI(1)	0.1,0.1, 0.5,0.5	0.1,0.1, 0.5,0.5	0.1,0.1, 0.5,0.5	0.1,0.1, 0.5,0.5	0.1,0.1, 0.5,0.5	0.1,0.1, 0.5,0.5	0.1,0.1, 0.5,0.5	0.1,0.1, 0.5,0.5	0.1,0.1, 0.5,0.5	0.1,0.1, 0.5,0.5	0.1,0.1, 0.5,0.5
	WFI(2)	0.2,0.5, 0.5,0.2	0.2,0.5, 0.5,0.2	0.2,0.5, 0.5,0.2	0.2,0.5, 0.5,0.2	0.2,0.5, 0.5,0.2	0.2,0.5, 0.5,0.2	0.2,0.5, 0.5,0.2	0.2,0.5, 0.5,0.2	0.2,0.5, 0.5,0.2	0.2,0.5, 0.5,0.2	0.2,0.5, 0.5,0.2
	WFI(3)	~	~	~	~	~	~	~	~	~	~	~
CIF	CIF(1)	0.9999	0.9999	0.9166	0.9166	0.9166	0.9999	0.9285	0.9999	0.7499	0.7499	0.7499
	CFI(2)	0.6428	0.8571	0.9285	0.9285	0.9285	0.9166	0.8571	0.6428	0.9999	0.9999	0.9999
	CFI(3)	~	~	~	~	~	~	~	~	~	~	~
GIC		0.9999	0.9999	0.9285	0.9285	0.9285	0.9999	0.9285	0.9999	0.9999	0.9999	0.9999
IC	Logical	L	L				L	L	L			
	Illogical			ILL	ILL	ILL				ILL	ILL	ILL
SOP		0.1,0.1, 0.2,0.3, 0.2,0.3, 0.2	0.2,0.3, 0.2,0.3, 0.2,0.3, 0.1	0.4,0.4	0.4,0.4	0.4,0.4	0.4,0.2, 0.4,0.1, 0.1,0.1, 0.1	0.2,0.4, 0.2,0.4, 0.2,0.4, 0.1	0.1,0.1, 0.2,0.3, 0.2,0.3, 0.2	0.2,0.2	0.2,0.2	0.2,0.2

A=ADENINE,T=THYMINE,G=GUANINE,C=CYTOSINE,U=URACIL,PPO=PREPROCESSOR OUTPUT,WFI=WEIGHT FOR INFERENCE,CIF=CATEGORY INFERENCE FUNCTION, GIC=GREATEST INFERRED CATEGORY,IC=INFERRED CATEGORY,L=LOGICAL,ILL=ILLOGICAL,SOP=SEPARATOR OUTPUT

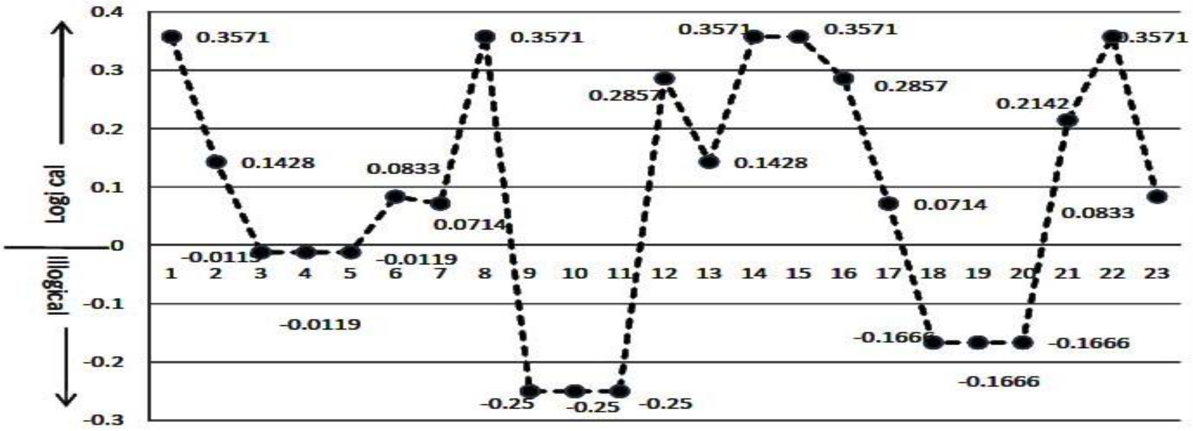


Figure 4 Chart showing logical sequence and illogical sequence from DNA Input

If CIF (1) > CIF (2) the inferred category is logical else if CIF (1) < CIF (2) then inferred category is illogical and the difference between CIF (1) and CIF (2) is plotted using graph shown in figure 4.

5. IDENTIFICATION OF DNA SAMPLES

SAMPLE 1 HUMAN-1 with BASE PAIR =32 and SEQUENCE =25
AATGTGTTGTGTGACCCCTCAAAATCTCTCAAATGTG
TTTTTACACTCCGTTGGTAATATGGAATGTGTTAAAGT
TGCTACCCGGGGTTTTTTAATGTGTCTCT

SAMPLE 2 HUMAN-2 with BASE PAIR =37 and SEQUENCE =30
CAAGTGTGTGGTTACCAAAATCTCTCAAATGTGGTGG
TTGGGCGTGGTTAAATATGGTAATGTGTTAAAGTGGT
GGTTTGTGGTTAGGGGGGGGCG

For DNA inputs from HUMAN 1 whose category is logical the corresponding seven consecutive components in the DNA sample is chosen as single sequence with base pair thirty two.

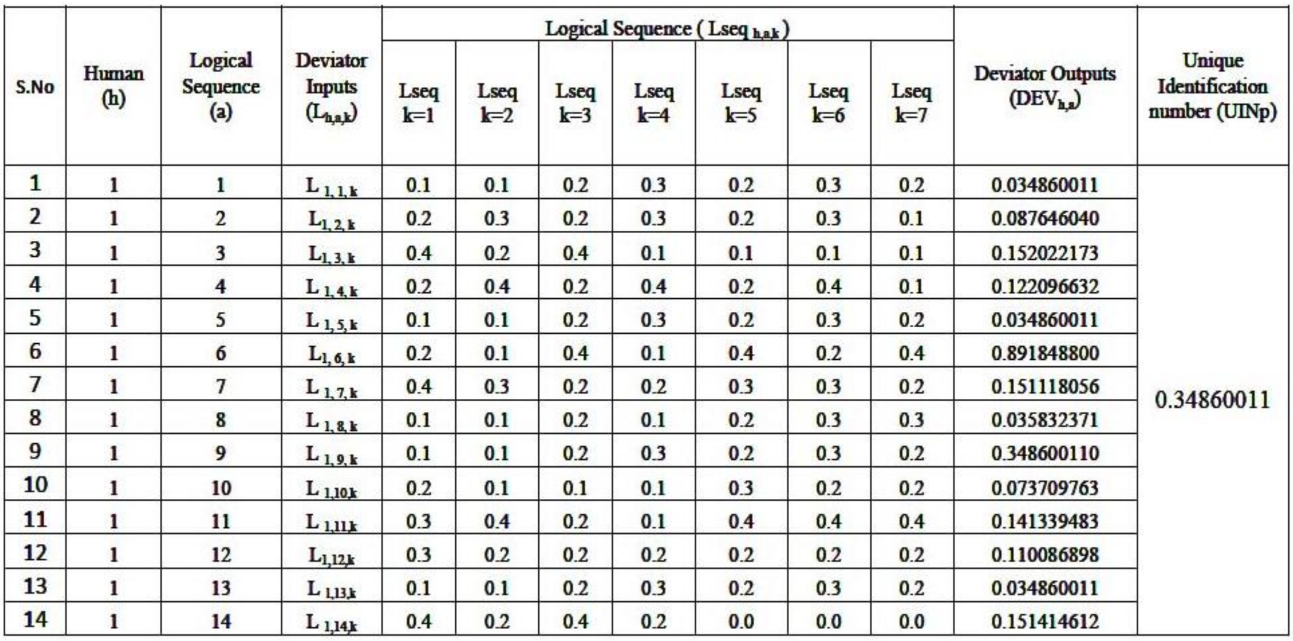
The DNA inputs whose category is illogical, the input is considered as a sequence with two consecutive components. To generate the unique identification number for each individual the set of logical sequences is used.

Logical Sequences (L) is given by

$$L_{p,\,s,\,k} = \text{Lseq}_{\ p,\,s,1}, \text{Lseq}_{\ p,\,s,2}, \ldots, \text{Lseq}_{\ p,\,s,\,k} \qquad \text{(Eq.7)}$$

$$\text{Where } \ p,s = 1 \text{ to } \alpha$$
$$\text{and } \quad k = 1 \text{ to } 7$$

Table 3 Deviator outputs for various logical sequence of Human-1



S.No	Human (h)	Logical Sequence (a)	Deviator Inputs ($L_{h,a,k}$)	Logical Sequence ($\text{Lseq}_{h,a,k}$)							Deviator Outputs ($DEV_{h,a}$)	Unique Identification number (UINp)
				Lseq k=1	Lseq k=2	Lseq k=3	Lseq k=4	Lseq k=5	Lseq k=6	Lseq k=7		
1	1	1	$L_{1,1,k}$	0.1	0.1	0.2	0.3	0.2	0.3	0.2	0.034860011	0.34860011
2	1	2	$L_{1,2,k}$	0.2	0.3	0.2	0.3	0.2	0.3	0.1	0.087646040	
3	1	3	$L_{1,3,k}$	0.4	0.2	0.4	0.1	0.1	0.1	0.1	0.152022173	
4	1	4	$L_{1,4,k}$	0.2	0.4	0.2	0.4	0.2	0.4	0.1	0.122096632	
5	1	5	$L_{1,5,k}$	0.1	0.1	0.2	0.3	0.2	0.3	0.2	0.034860011	
6	1	6	$L_{1,6,k}$	0.2	0.1	0.4	0.1	0.4	0.2	0.4	0.891848800	
7	1	7	$L_{1,7,k}$	0.4	0.3	0.2	0.2	0.3	0.3	0.2	0.151118056	
8	1	8	$L_{1,8,k}$	0.1	0.1	0.2	0.1	0.2	0.3	0.3	0.035832371	
9	1	9	$L_{1,9,k}$	0.1	0.1	0.2	0.3	0.2	0.3	0.2	0.348600110	
10	1	10	$L_{1,10,k}$	0.2	0.1	0.1	0.1	0.3	0.2	0.2	0.073709763	
11	1	11	$L_{1,11,k}$	0.3	0.4	0.2	0.1	0.4	0.4	0.4	0.141339483	
12	1	12	$L_{1,12,k}$	0.3	0.2	0.2	0.2	0.2	0.2	0.2	0.110086898	
13	1	13	$L_{1,13,k}$	0.1	0.1	0.2	0.3	0.2	0.3	0.2	0.034860011	
14	1	14	$L_{1,14,k}$	0.4	0.2	0.4	0.2	0.0	0.0	0.0	0.151414612	

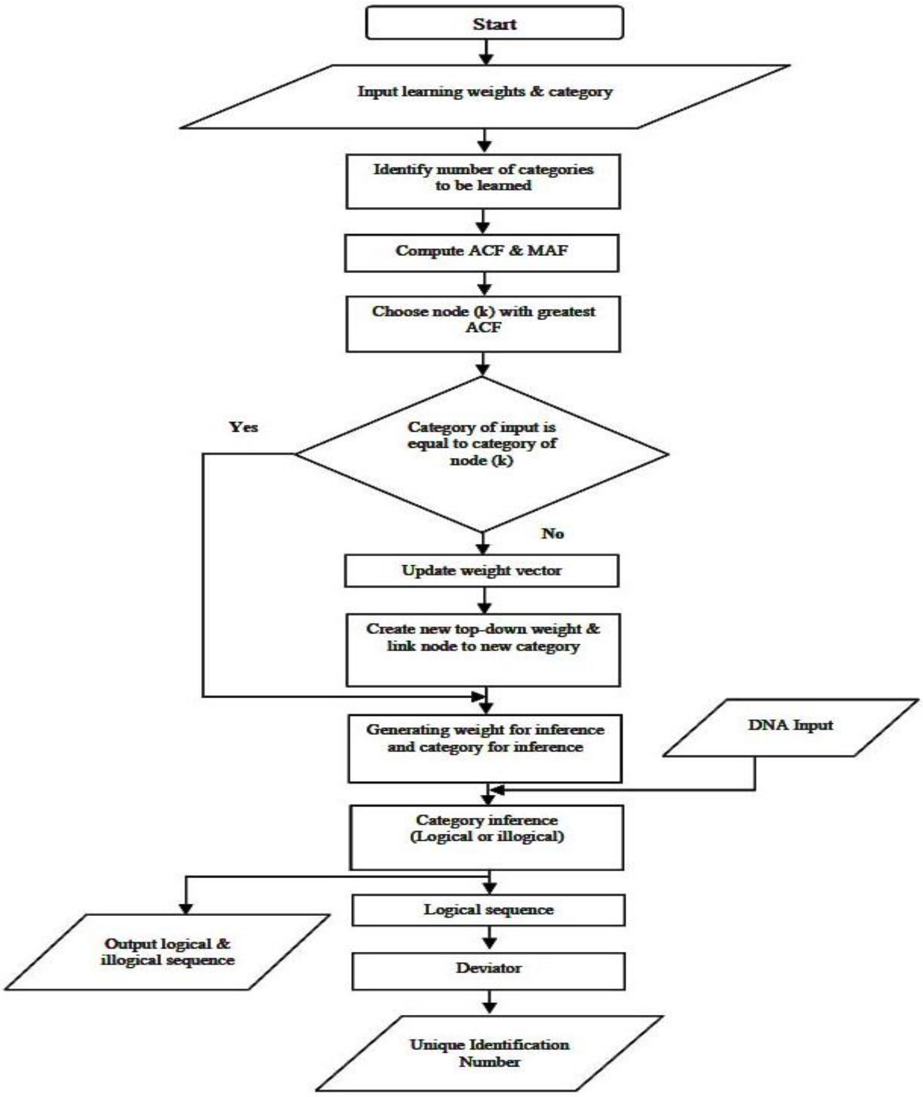

Start
Input learning weights & category
Identify number of categories to be learned
Compute ACF & MAF
Choose node (k) with greatest ACF
Category of input is equal to category of node (k)
Yes
No
Update weight vector
Create new top-down weight & link node to new category
Generating weight for inference and category for inference
DNA Input
Category inference (Logical or illogical)
Logical sequence
Output logical & illogical sequence
Deviator
Unique Identification Number

Figure 5 Flow chart showing NFPRS System

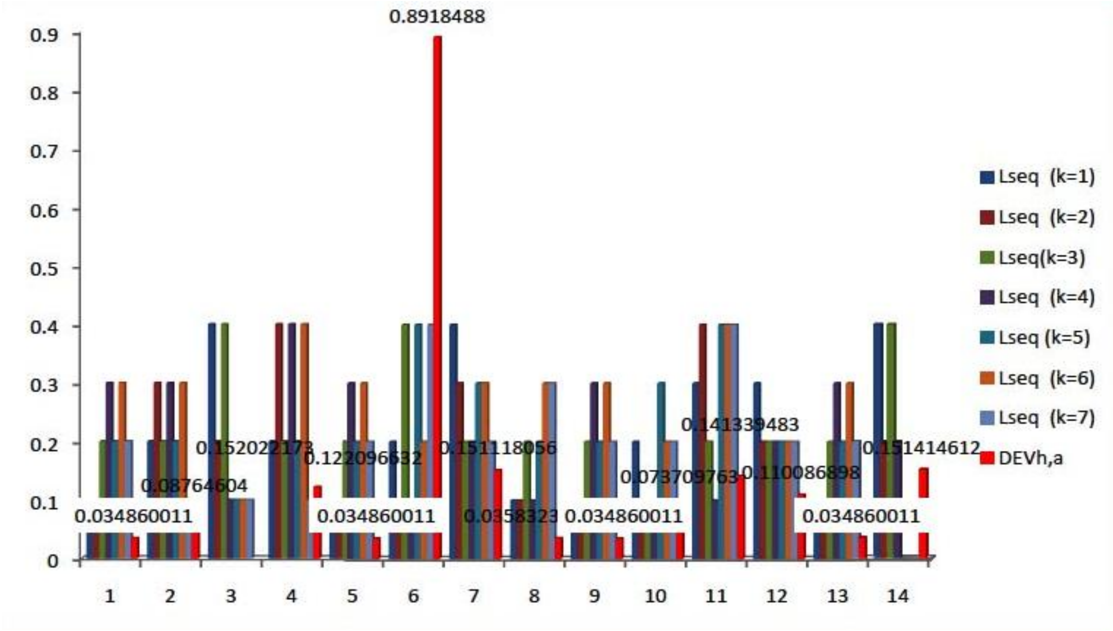


Figure 6 Chart shows the unique identification number in deviatior outputs

The mean of the logical sequence is given by

$$\overline{x}_{h,a} = \frac{\sum_{k=1}^{7} k\,(Lseq_{h,a,k})^{k}}{m} \qquad \text{(Eq.8)}$$

Where $h, a = 1$ to α

and $m = 7$

The variance of the logical sequence is calculated from the mean of the logical sequence is given by

$$Var_{h,a} = \frac{\sum_{k=1}^{7}\left[k(Lseq_{h,a,k})^{k} - \overline{x}_{h,a}\right]^{2}}{m-1} \qquad \text{(Eq.9)}$$

Where $h, a = 1$ to α

and $m = 7$

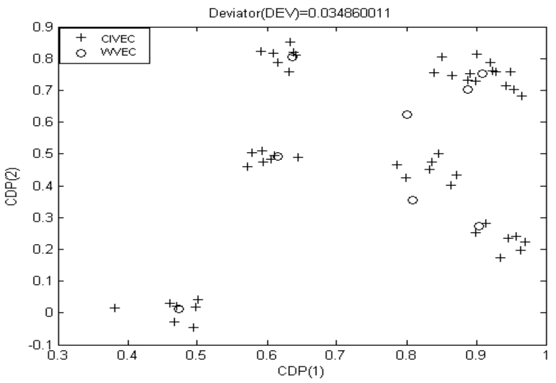
The deviation of Human-1 calculated by

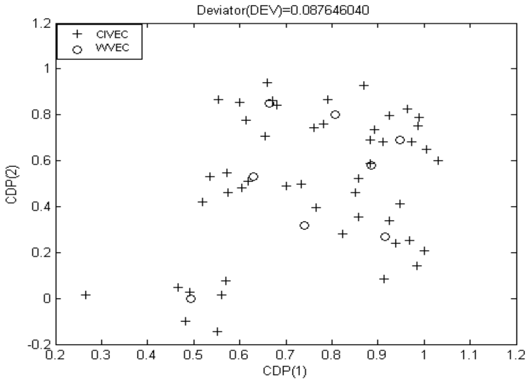
$$Dev_{h,a} = \sqrt{Var_{h,a}} \qquad \text{(Eq.10)}$$

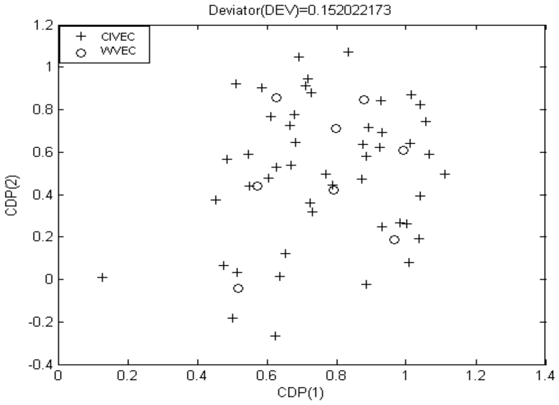
Where $h, a = 1$ to α

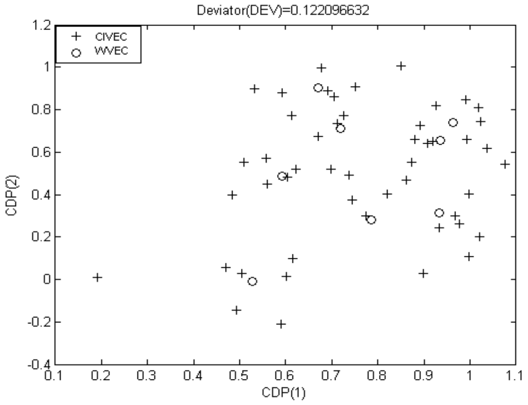
generates the unique identification number and is shown in figure 6. The deviation of the repeated logical sequences is calculated and plotted using MATLAB. Figure 7 explains that the clusters formed for repeated sequences are unique.

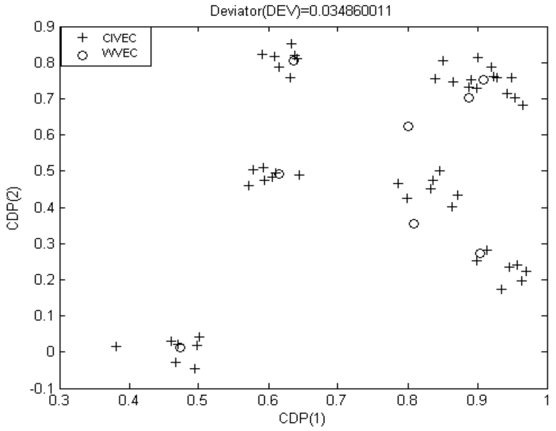
6. MATLAB OUTPUT

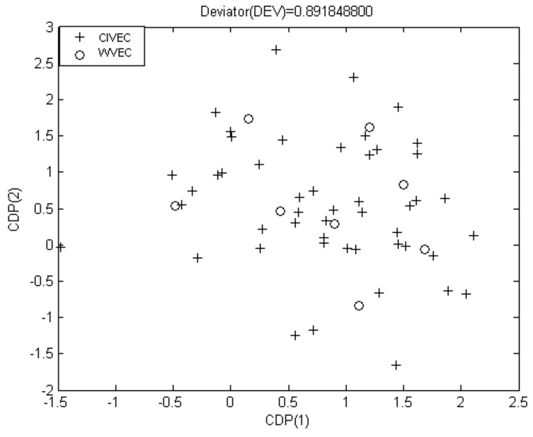


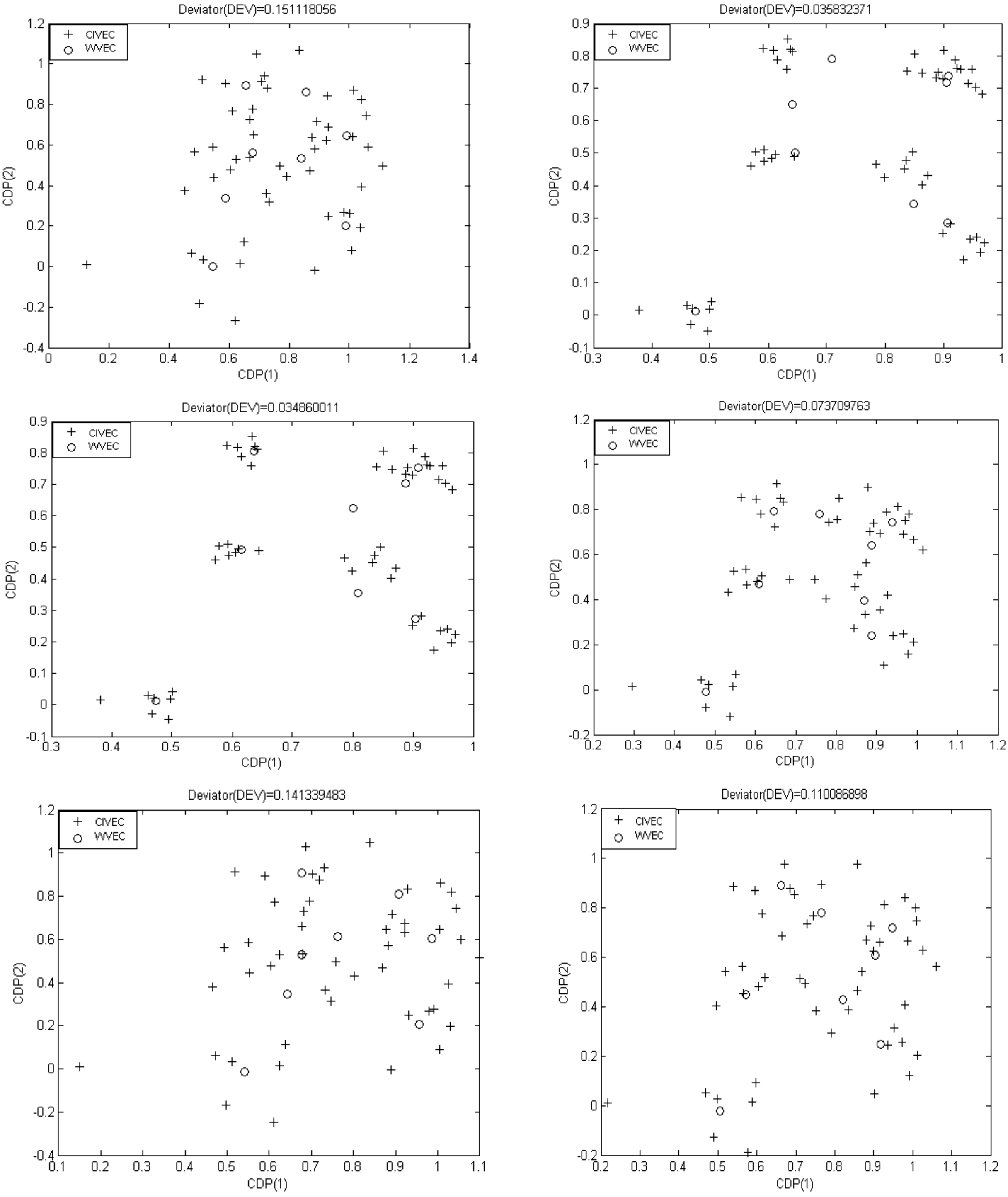












CDP = Clustered Data Points, CIVEC= Cluster of Input Vectors, WVEC=Weight Vectors

Figure 7 Cluster generated for deviator outputs

7. CONCLUSION

As an attempt to the automation of the DNA finger printing, the Neural-Fuzzy Pattern Recognition system discussed in the above work is used to classify the sequences that are used to identify a unique number from the given human DNA sample. Also in the field of Bio-informatics it would be inevitable in future to deal with numerous nucleotides compositions, their classifications in case of protein sequencing, mutation and so on, for which the above tool can be used extensively since it uses neural and fuzzy approach with a long range of fuzzy values between 0 to 1 instead of a minimal range of real values. The NFPRS system also overcomes complications in generating genetic algorithms that require more iteration when conventional mathematical techniques are applied.

8. ACKNOWLEDGEMENT

The authors would like to thank Dr.K.Somasundaram at Amrita University, Coimbatore and Dr.N.Nagaveni at Coimbatore Institute of Technology, Coimbatore for their help in conducting the experiments.

9. REFERNCES

[1] Carpenter, G.A. and S. Gross berg, and David B. Rosen, “ARTMAP: Supervised Real-time Learning and Classification of Non-stationary Data by a Self-organizing Neural Network”, Neural Networks, Vol 4, pp.565-588.

[2] Schalkoff, Robert (1992), “Pattern Recognition- Statistical, Structural and Neural Approaches”, John Wiley & Sons.Timothy, J. Ross (1997), “Fuzzy Logic with Engineering Applications”, McGraw Hill, 1997.

[3] Freeman,J.A. & D.M.Skapura(1991), “*Neural Networks*”, Addison Wesley.

[4] Carpenter, G.A. and S. Gross berg, Natalya Markuzon, J.H. Reynolds, and D.B.Rosen(1992), FuzzyARTMAP: “A Neural Network Architecture for Incremental Supervised Learning of Analog Multi-dimensional Maps”, IEEE Trans on Neural Networks, 1992, Vol 3, No. 5 pp. 698-713.

[5] kasubam,Tom (1993), “Simplified Fuzzy ARTMAP, AI Expert” , November, pp. 18-25 and Pao, Y.H.(1989), “Adaptive Pattern Recognition Neural Networks”, Addision Wesley, M.A.

[6] Baxevanis, A.D. and Ouellette, B.F.F., eds., “Bioinformatics: a Practical Guide to the Analysis of Genes and Proteins”, third edition. Wiley2005.

[7] AchuthsankarS Nair Computational Biology & Bioinformatics - A gentle Overview. “Communications of Computer Society of India”, January 2007 and Pevzner, Pavel A. “Computational Molecular Biology: An Algorithmic Approach” The MIT Press, 2000.Zhang, Z.,Cheung, K.H. and Townsend, J.P.Bringing Web2.0 to bioinformatics.

[8] Hogeweg, P. and B. Hesper B. (1984) “The alignment of sets of sequences and the construction of phyletic trees: an integrated method”. J Mol Evol 20: 175-186.

[9] Saenger, Wolfram (1984). “Principles of Nucleic Acid Structure” New York: Springer-Verlag. Butler, John M (2001) “Forensic DNA Typing” Elsevier. pp. 14–15. Pearson H 2006. "Genetics: what is a gene?” pp.398–401.

[10] Wolfsberg T, McEntyre J, Schuler G (2001). “Guide to the draft human genome”. Nature 409 (6822): 824–6.and “Designation of the two strands of DNA” JCBN/NC-IUB Newsletter 1989, Accessed 07 May 2008.

[11] Nature Archives Double Helix of DNA: 50 Years effreys A, Wilson V, Thein S (1985). “Individual-specific 'fingerprints' of human DNA”.

[12] Effreys A.J., Wilson V., Thein S.W 1984. “Hyper variable minisatellite regions in human DNA” Nature 314: 67-73

[13] Joseph Wambaugh “The Blooding” New York: A Perigord Press Book, 1989, 202, Rose & Goos.

[14] "DNA - A Practical Guide" (Carswell Publications, Toronto) and Bhattacharya, Shaoni (20 April 2004). “Killer convicted thanks to relative's DNA".

[15] Min JouW, HaegemanG, Ysebaert M, Fiers W May 1972 “Nucleotide sequence of the gene coding for the bacteriophage MS2 coat protein”. Fiers W, Contreras R,Duerinck F, (April 1976).

[16] “Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene”.Gilbert, W. “DNA Sequencing and Gene Structure”. Nobel lecture, 8 December 1980 and Sanger F. “Determination of nucleotide sequences in DNA”. Nobel lecture, 8 December 1980.

[17] Lipman, DJ, Pearson, WR (1985) “Rapid and sensitive protein similarity searches”. Science **227** (4693): 1435–41 and Pearson, WR; Lipman, DJ (1988) “Improved tool for Biological Sequence Comparison”. Proceedings of the National Academy of Sciences of the United States of America 85 (8):2444–8.

[18] Wong, P.M., I.J. Taggart, and T.D. Gedeon(1995), “The Use of Fuzzy ARTMAP for ithofacies Classification: Comparison Study”, Proc. Of SPWLA, Thirty sixth Annual Logging Symposium, Paris.

[19] Barbara Moore “ART1 and Pattern clustering” In David Toruretzky, Geoffery Hinton, and Terrence Sejnowski, editors, Proceedings of the 1988 Connectionist Models Summer School, pages 174-185.