Sri Lanka Association for Artificial Intelligence (SLAAI)
Proceedings of second Annual Session
19th February 2005 – Colombo

"Gene Doctor"

System that Diagnoses Genetic Diseases by Analyzing DNA Sequences Through use of "Artificial Neural Networks".

A.J. Peiris
Informatics Institute of Technology
57 Ramakrishna Rd, Colombo 06, Sri Lanka.
janaka2000@gmail.com

A.Wickramasinghe
Informatics Institute of Technology
57 Ramakrishna Rd, Colombo 06, Sri Lanka.
a.wickramasinghe@iics.ac.lk

Abstract

Humans are at times infected with genetic diseases caused by a mutation in DNA. Getting a DNA test done is a very strenuous, time-consuming and costly task, especially due to the complexity, and the unpredictable nature of mutations. Therefore, once a human get infected with a genetic disease, he or she has to go through a lot of hassles before getting a good diagnosis for his or her sickness. Artificial intelligence can be used to solve the above-mentioned problem. There are many artificial neural network techniques that could be used to solve this kind of problems. But, artificial neural networks would be the best artificial intelligence technique to solve this particular problem due to the complexity and unpredictable nature of the DNA sequence." Gene Doctor" is a computer system that diagnoses genetic diseases through analysis of DNA sequences. The analyzing of the DNA sequence is done through an artificial neural network. The artificial neural network used in " gene doctor" is a three layer neural network with Backpropagation as its Learning Rule, for the working prototype of "Gene Doctor". The prototype is currently trained to detect four most common genetic diseases with an accuracy of 78%."Gene Doctor" also includes some other features such as Gene Therapy Finder, Graphical Sequence Simulator, Detail DNA Report, and Automated Translation & Transcription. The above features are helpful to doctors and scientists and lab assistance that work in microbiology and DNA research field.

1. Introduction

Many millions suffer from genetic diseases. "Gene Doctor" has been implemented with the

aim to provide quick and efficient solution and as enhanced service inside a microbiology laboratory, with the help of artificial intelligent techniques.

Five features have been implemented in "Gene Doctor" to assist doctors and research staff in the field of microbiology.

The features are as follows:

- ***Disease Detector*** - Diagnose genetic diseases by analyzing patient's DNA Sequence.
- ***Therapy Finder*** - Recommendation of Gene therapy for genetic diseases.
- ***Sequence Simulator*** - Graphically display a given DNA Sequence.
- ***DNA Report*** - Produce a report of a given DNA sequence
- ***DNA -> RNA -> Protein*** – Generate RNA & Protein sequences for a given DNA sequence (Translation & Transcription).

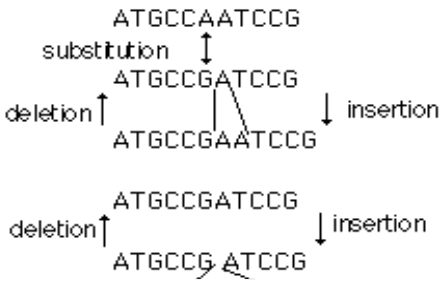
The above functionality has been obtained by using several techniques such as artificial neural networks, database transactions, and general algorithms along with multimedia.

The main artificial intelligence technique which is used in "Gene Doctor" is artificial neural networks. Fuzzy logic and genetic

algorithms are also used to quicken the artificial neural network training process.

2. Genetic Disease Detection

Most of the genetic diseases are caused by *mutations* (differences in the DNA of an individual as compared to normal human DNA). Sickle cell anemia, a recessive disorder is a relatively simple example for a genetic diseases caused by mutation. The type of mutation exemplified in sickle cell anemia is called a substitution, because one nucleotide base is substituted for another. Other types of mutations include insertions and deletions.



Types of mutation

Figure 1

Figure 1 illustrates the types of mutation that can occur which are subtraction, insertion and deletion.

DNA testing and *gene mapping* are two methods of detecting mutations in a DNA sequence. Among the two, the most commonly used is DNA testing. Two different types of DNA tests exist, namely,

1. ***RFLP*** (Restriction fragment length polymorphism)
2. ***PCR*** (Polymerize chain reaction)

PCR is the most widely and commonly used technique in genetic disease detection.

Presently, the genetic disease detection is done manually by experts who analyze output of the PCR and determine whether or not a particular DNA Sequence carries a genetic disease.

Drawbacks of existing genetic diseases diagnosis process

1. *Time consuming*
2. *Costly*
3. *Exposed to human error*
4. *Lack of experts*

(Especially in developing countries)

"Gene Doctor" provides a solution to the above-mentioned drawbacks by automating the task performed by an expert with the use of Artificial Neural Networks.

3. Use of Artificial Neural Network

Artificial Neural Networks have been applied to solve many pattern recognition and classification problems in the field of biology. Though expert biologists typically achieve higher levels of accuracy in heuristic predictions, several neural network-based methods have eventually contributed significantly in advancing the field of bio-informatics, and some are clearly influencing molecular biology. Therefore, this technology was implemented in "Gene Doctor"

During the implementation of "Gene Doctor", selecting the suitable topology and training the neural network was a very tough and demanding task, especially due to the large amount of data embedded in a single DNA.

Many support tools were implemented in order to achieve maximum results in training the Artificial Neural network. The DNA sequence, which is represented in a four base code, was used as the input to the network. The DNA sequence that is represented by a formula containing the English letters A, C, G, and T was formatted digitalized before use.

3.1 Inputs for the ANN

DNA sequences taken out from the DNA report was used as the input for the Artificial Neural network.

The DNA sequence is represented by a formula containing a four base code (A, C, G, and T).

This code was digitalized and used as the input to the Artificial Neural Network.

A (Adenine) = 1,1,1,
C (Cytosine) = 1,0,1,
G (Guanine) = 0,1,1,
T (Thymine) = 1,0,1,

Adenine, Cytosine, Guanine and Thiamine are the names of the four bases, which a DNA sequence is made of.

3.1.1 Formatting Input

Considerable amount of cleaning were needed to be done to the inputs before it was digitalized.

Once a DNA sequence is generated using RFLP or PCR the DNA sequence is formatted for a more human readable format.

Few tools were used in "Gene Doctor" to format the Input and digitalize it. The tools are as follows:

- *Sequence Cleaner*
- *Binary Converter*

Since the DNA sequence are in different lengths and some DNA sequences could be extremely lengthy, all DNA sequence were assigned a standard length before in was digitalized.

Initially a very short length of the DNA sequence was used to make the training presses of the artificial neural network faster.

If the length of the DNA sequence was shorter than the predefined length, a set of zeros were added to the end of the digitalized until the required length is obtained.

E (Extra Digit) = 0,0,0,

Once the DNA sequence was formatted, digitalized and standardized length was sent as the input for the artificial neural network.

3.2 Topology of the ANN

Lengthy experiments were carried out with different combinations and the topology with the following combination was selected as the most suitable for extensive training.

Learning Rule : Backpropagation
Activation Function : Bipolar Continues
Number of Layers : 3
Number of Components : 15000
Number of nodes in the Input Layer : 2
Number of nodes in the Hidden Laye : 12
Number of nodes in the Output Laye : 3
Neta : 0.01
Lambda : 1
Alpha : 1

Most of the parameters used for the topology were obtained trough extensive training of the artificial neural network.

Tools such as topology tester and ANN trainer were used to get the best combination for the artificial neural network.

The literature survey was also extremely useful to obtain the best combination for the topology

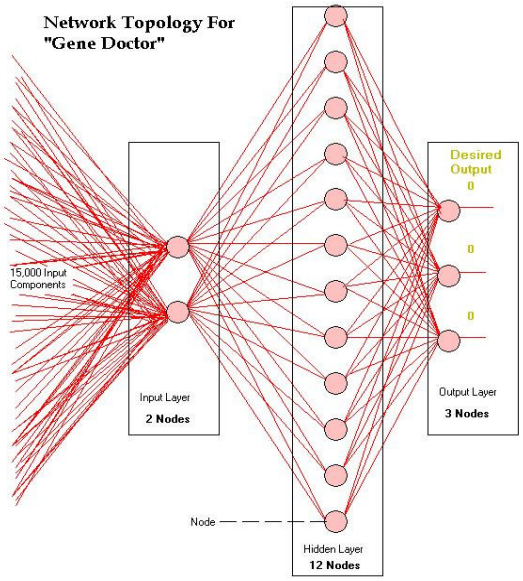


Figure 2 graphically illustrates topology of the Artificial Neural Network used in "Gene Doctor".

Long hours of training were put in to the artificial neural network in order to achieve a higher level of accuracy.

3.2.1 Training the ANN

The artificial neutral network was trained using the tool "ANN Trainer". This was the most complicated screen and was the heart of "Gene Doctor".

There were few features inbuilt to this screen where the user could

- *Format inputs*
- *Save digitalized inputs*
- *Generate new sets of weights*
- *Save sets of Weights*

A genetic algorithm is used to generate weights at this point. This will make the training process much faster and efficient.

There are some other features such as momentum implemented so that the training of the artificial neural network will be faster and

will prevent the artificial neural network from paralyzing.

Once the network is trained to a certain extent the weights can de saved and then assigned to the final artificial neural network used by "Disease Detector" and "Disease Detector Express".

3.3 Outputs of the ANN

The system was initially trained only to detect four most common diseases.

The four diseases which were captured through the output layer by digitizing are as follows:-

Siecle Cellanigme	*- 1,1,1*
Homo Cystinuria	*- 1,1,0*
Thalassemia	*- 1,0,1*
Galactosemia	*- 0,1,1*

The artificial neural network could be trained to detect seven diseases using above topology.

The neural network can be even trained for more diseases as a future enhancement.

4. Gene Therapy

Gene therapy uses the technology of genetic engineering to cure or treat a disease caused by a gene that has changed in some way (mutated).

One method is replacing sick genes with healthy ones. Gene therapy trials and other research may lead to new ways to treat or even prevent many diseases.

"Gene Doctor" provides the functionality of recommending gene therapy for genetic diseases.

This is one of the five features facilitated in "Gene Doctor".

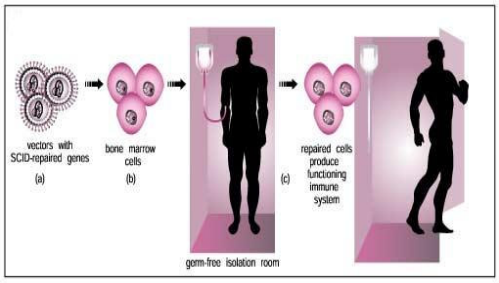


Figure 3 is a simple illustration on how Gene Therapy is used to cure genetic diseases.

"Gene doctor" uses general database transactions to recommend gene therapy.

Once, someone has been diagnosed with a genetic disease, "Gene doctor" has the ability to recommend gene therapy according to the illness.

4.1 Typical Gene therapy process

1. To reverse disease caused by genetic damage, researchers isolate normal DNA and package it into a vector (a molecular delivery truck usually made from a disabled virus).

2. Doctors then infect a target cell (usually from a tissue affected by the illness, such as liver or lung cells) with the vector.

3. The vector unloads its DNA cargo, which then
begins producing the missing protein and restores
the cell to normal

5. DNA -> RNA -> Protein Conversion

"Gene Doctor" has the feature to Generate RNA & Protein sequences for a given DNA sequence. This conversion is done in two steps. They are

Translation : DNA to RNA Conversion
Transcription : RNA to Protein Conversion

There are variety types of RNA. (Among the types of RNA, some of them are messenger RNA or mRNA, ribosomal RNA or rRNA and transfer RNA or tRNA.)

When a protein is to be produced, the DNA separates and a copy of the gene is transcribed (the process is referred to as transcription by molecular biologists) into a strand of RNA.

In this process, converting RNA sections from the original DNA are known as introns.

But, these are not involved in the production of a protein, are separated or snipped out and a resulting string of exons from the original DNA is left.

What is left is sometimes referred to as a coding sequence, and it contains the codons (in the letters A, C, G, and U)

And finally, this will create the protein, by the process usually referred to as translation.

This process is automated in "Gene Doctor". Sets of Algorithms are used to automate this process.

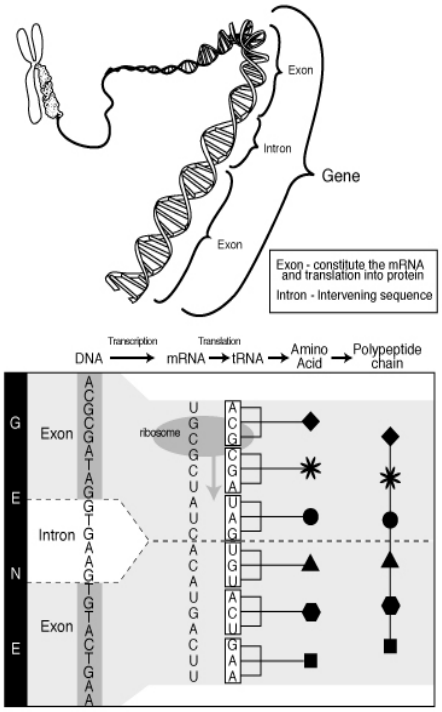


Figure 4

Figure 4 illustrates the translation and transcription processes.

6. Other features in Gene Doctor

Other than first three features, which have been discussed in detail, two more additional features are implemented in "Gene Doctor" These features are as follows:

6.1 Sequence Simulator

Graphically display a given DNA Sequence. "Gene Doctor" has the ability to graphically simulate any given DNA with the use of dynamically created imagers

6.2 DNA Report

"Gene Doctor" Produces a report of a given DNA sequence. This printable report contains information such as:

Species
This field specifies the species of the DNA Sequence owner. "Gene doctor" only uses human (Homo Sapiens) DNA. But the species is mentioned on to of the DNA report since it is a kind of a tradition. And it will also be useful because Animal DNA Analysis will be implemented in "Gene Doctor" as a future enhancement.

Gene:
This field states the gene of the DNA sequence. Though we call the sequence as DNA Sequence, we only consider the specific gene in the DNA Sequence. This is because the DNA sequence is too large and we only consider the gene, which is the source of the genetic diseases. This issue will be discussed in detail in the "problems encountered" section.

Number of Base pairs:
This is the Length of the sequence

Sequence:
The DNA sequence of particular species

7. Problems Encountered

7.1Genetic Disease due to Chromosomes

During the literature survey, it was found that genetic diseases are also caused by change in the number of chromosomes. Usually the human body carries a specific number of copies of each chromosome. But, genetic diseases can be caused by possession of too many or not enough copies of chromosomes. Detecting such diseases could not be done from the proposed methodology.

Since Genetic Disease caused due to Chromosomes is very rear, "Gene Doctor" was designed only to detect genetic diseases caused by mutation in the genetic sequence.

7.2 Largeness of the input sequence

Length of the DNA sequence was one of the major problems encountered. Since DNA sequence is the input for the neural network, it had to be in a manageable length as well as

complete enough to generate results. Two remedies were suggested.
Use only the sequence of the problematic area in the sequence as the input. Mutations of a particular genetic disease originate from a particular area of a DNA sequence.

By only using the particular area of the sequence as the input, same results can be obtained.

Other option is to define a large but finite length for a sequence, where the majority of the DNA sequences will fit in.
While carrying out the implementation it was found that the most suitable solution was the combination of the both solutions mentioned above.
After implementing the above solution the artificial neural network was working well.

But as the training require more and more resources (Processing power and memory). This resulted in system and memory errors.

And finally the maximum length of a DNA sequence was also cut down so that the artificial neural network will need fewer resources.

7.3 Training

High consumption of computer resources and network being paralyzed were the two main problems experience during the training process.

"Momentum" and generating weights through genetic algorithms were implemented to address the issues up to a certain extent.

The implementations were done according to literature surveys carried out.

7.4 Graphical display of DNA sequence

Large length of the DNA sequence again generated problems when trying to display the sequence in "Sequence Simulator".

Visual basic only allowed a limited number of controls to be created in a single VB form. But to graphically create one four-base out of the DNA sequence out of dynamically created controls, it required 13 controls.

Once calculated the number of controls needed to be dynamically created to simulate one DNA sequence summed up to 16,250.

Therefore, finally imagers were created of all 256 combinations of the four-base pairs and then loaded them in to visual basic form so that a DNA sequence could be graphically simulated.

This enables to display a DNA sequence over of a length close to the maximum length of DNA sequence used in "Gene doctor.

7.5 Gene therapy

To develop the functionality of recommending Gene therapy for the genetic disease it was planned to use an expert system.

But with the data collected, it was found that finding the gene therapy for a known disease was quite straightforward.

Therefore this feature was implemented using SQL statements.

8. Future Enhancements

"Gene Doctor" is only a working proto type. There for, modifications and enhancements are needed to implement it as a commercial product.

Below are some future enhancements that could be considered.

The concept of gene doctor that is to analyze DNA sequence using artificial neural networks can be used in many areas in DNA research.

For example -
Evolution pattern detection
Drug resistance (Human/Anima)
Prehistoric species recognition
Forensic Investigations

DNA sample testing machines can also be plugged, to make a complete product in genetic disease diagnosis.
Diagnose genetic disease caused from both mutations and chromosomes.
The neural network can be trained to diagnose more genetic diseases.

9. Conclusion

The heart of "Gene Doctor" is the manner in which "Artificial Neural Networks" can be used to recognize patterns in a DNA sequence; along with how multipurpose software could support and enhance functions in a DNA laboratory or in a research facility in countries such as Sri Lanka.

The concept of pattern recognition in DNA sequences can be used extensively in many other subject areas in DNA research.

Evolution pattern detection, Drug resistance (Human/Anima), Prehistoric species recognition, Forensic Investigations are among some of the examples.

More training of the artificial neural network involving more time processing power and larger number of test inputs could significantly enhance accuracy of the system.

Once a high level of accuracy is achieved "Gene Doctor" the computer system that diagnoses genetic diseases, could be implemented commercially.

10. References

[1] Zurada, M. Jacek (1992) Introduction to Artificial Neural Systems : West publishing company

[2] Carpenter, G.A and Grossberg, S.(1991) Pattern recognition by self organizing neural networks. Cambridge MA: MIT Press.

[3] Pamala,C. and Richard, A.(1994) Bioche mistry second edition : Lippincott – Raven.