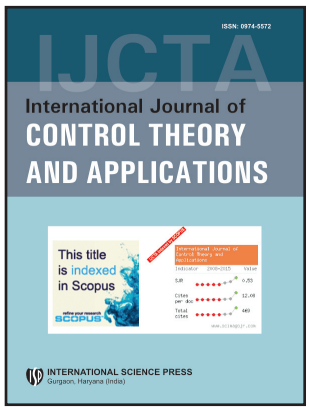

International Journal of Control Theory and Applications

ISSN : 0974-5572

Volume 9 • Number 43 • 2016

Disease Detection Using DNA-RNA Patterns

Lidiya Nixon[a] and Joby George[b]

[a-b]Department of Computer Science and Engineering, M.A College of Engineering, Kothamangalam, Kerala, India. Email: [a]lidiya.nixon@gmail.com; [b]jobygeo@hotmail.com

Abstract: The proposed system basically aims at analyzing the DNA structure of the host in comparison with that of the causative organism. It basically consists of 4 modules. Initially the validity of data sample is checked. This checks for the presence of valid bases within the obtained sample. This can identify the presence of invalid mutations whether present or not within the new sample introduced. The pattern recognizing mechanism makes use of sampling the whole sequence and then generating the corresponding hash key pair. Once the validity is checked then comes the sequencing and sampling of the DNA data of the sample. This is done using the basic DNA Codon. The next involves processing the incoming (disease causing) DNA there by analyzing its pattern and mutation percentage. Also the novel algorithm proposes a way to match the host DNA with that of the incoming one. This is done by evaluating the structural similarity, sequence similarity as well as other important basic DNA parameters. The next step involves the prediction of disease that the host will be getting in near future. This involves the modification of the algorithm to the next level that makes it predict the disease. Further a different causative organism can come in future which has just few variations in the genetic composition as compared to the current organism. This few gene variations can be predicted using genetic algorithm. This novel method can give the disease prediction more accuracy. Also the mechanism will show the percentage of correction that the visitor DNA had done to the host DNA.

Keywords: Disease Detection; DNA; RNA; Pattern Matching.

1. INTRODUCTION

A disease as defined by etiology, the study of causation of a disease, is an abnormal condition of a part, organ, or system of an organism resulting from various causes. Disease is often considered as a medical condition in association with specific symptoms and signs. The cause can be external factors such as pathogens, or internal dysfunctions particularly of the immune system such as an immunodeficiency, or a hypersensitivity including allergies and autoimmunity.

Nucleic acids are of two types namely Deoxy Ribo Nucleic Acid (DNA) and Ribo Nucleic Acid (RNA). DNA is constantly subject to mutations which are accidental changes in its code. Mutations can lead to missing or malformed proteins, and further lead to disease. Normally we all start out our lives with some mutations which are inherited from our parents called germ-line mutations. However, mutations can also be acquired during our

lifetime. Some mutations happen during cell division which is the process where DNA gets duplicated. Still other mutations are caused when DNA can get damaged by environmental factors, including UV radiation, chemicals, and viruses. Some mutations are harmful while some others are beneficial.

Today in this modern world we hear daily the emergence of new diseases. The next thing that bothers us on being aware of such a new disease will be whether we will get affected by it or not. That remains as a threat for us until its main cause as well as mode of transmission is identified. Here a system is thus introduced to identify or predict the chance of occurrence of a new disease that comes into the arena.

The objective of our approach is to develop a prediction system. Its goal is to predict the possibility of occurrence of a disease within a particular organismic group when the genetic constitution of both the causative organism and the host that may get affected by the organism is known. For this we have to identify the genetic match between the causative organism and the host. And if a desired match level is obtained upon the evaluation then there exists a possibility that the particular organism may get affected by the disease.

Identification of basic cause as well as the chance of occurrence of a new disease whenever a new disease comes into the arena has become a greatest challenge. Much of the focus of human disease genetics aims at identifying nucleotide variants that contribute to disease phenotypes. This is a complex problem, often involving contributions from multiple loci and their interactions, as well as effects due to environmental factors.

The difference with the normally used approaches for disease detection is that here the prediction mechanism makes use of the basic genetic constitution rather than the phenotype features or external factors.

2. RELATED WORK

Recently the focus of human disease genetics is directed towards identifying nucleotide variants which contribute to disease phenotypes. This is a complex problem that involves contributions from multiple loci and their interactions along with effects due to environmental factors. Even though some diseases with a genetic basis are caused by nucleotide changes that alter an amino acid sequence in other cases disease risk is associated with altered gene regulation [1]. Understanding the basic causes of human disease is one of the most fundamental goals of modern medicine. There can be difference in individuals with respect to disease susceptibility, disease progression and effectiveness of treatment. The identification of factors contributing to these differences and then elucidating their interactions so as to decide their contribution to aspects of disease phenotype, is basically a precursor to improved prevention, detection as well as treatment of disease. The basic understanding of human disease derives from the study of those diseases that segregate in families in a Mendelian fashion, where the causative variants and the genes in which they reside have been identified through classical family linkage approaches and through studies in large pedigrees and in isolated populations based on founder effects. However the vast majority of common diseases exhibit a more complex mode of inheritance and while aggregating in families they rarely exhibit Mendelian inheritance.

Rapid progress in human genomic project has stimulated investigations for gene therapy and stimulated diagnosis of human diseases through mutation or polymorphism analysis of disease causing genes and has resulted in a new class of drugs. The recent development of capital electrophoresis technologies has facilitated the application of capillary electrophoresis to the analysis of DNA-based drugs and disease causing genes by capillary electrophoresis. Many successful applications of capital electrophoresis are the promising technology for DNA diagnosis of human diseases and quality control, pharmacokinetic analysis and therapeutic drug monitoring of DNA-based drugs.

The use of DNA analysis (by employing DNA probes) is a novel and revolutionary approach which specifically identifies the disease-causing pathogenic organisms. This is in contrast to the traditional methods of

disease diagnosis which is done by detection of enzymes, antibodies etc., besides the microscopic examination of pathogens. Although at present not in widespread use, DNA analysis may soon take over the traditional diagnostic tests in the coming years.

MicroRNAs (miRNAs) constitute a large family of noncoding RNAs that function as guide molecules in diverse gene silencing pathways. Current efforts are focused on them regulatory function of miRNAs, while not much is known about how these unusual genes themselves are regulated. Here the first direct evidence that miRNA genes are transcribed by RNA polymerase II (pol II) is presented. The primary miRNA transcripts (pri-miRNAs) mainly contains cap structures and poly(A) tails, which are the unique properties of class II gene transcripts[2]. This was thus considered an important discovery because most of the diseases are basically associated with miRNA genes.

Predicting miRNA genes is a challenging bioinformatics problem and normal computational methods have failed to deal with it efficiently. An approach was made by researchers that combined the efficiency of support vector machines and genetic algorithm in the embedded classification model. The embedded feature selection component selects a compact feature subset that contributes to the performance optimization. It gained better results compared to already existing methods [3].

Prediction of miRNA genes using machine learning approach was another important area of research. Most methods had drawback of high false positives or false negatives. One reason behind this was the deficiency of negative available samples. This made the training less efficient. Thus researchers made an attempt to include all available negative results available till then in the training [4].

Nearly 30 hereditary disorders in humans is the result of an increase in the number of copies of simple repeats in genomic DNA. These DNA repeats seem to be predisposed to such expansion because they have unusual structural features which can disrupt the cellular replication, repair and recombination machineries. The presence of expanded DNA repeats can even alter the gene expression in human cells, leading to disease. Surprisingly, many of these extenuating diseases are caused by repeat expansions in the non-coding regions of their resident genes. It is becoming clear through experimental observations that the peculiar structures of repeat-containing transcripts are at the heart of the pathogenesis of these diseases. It is generally thought that the unusual structural features of expandable repeats can predispose them to instability. Indeed, repeats that are not structure-prone seems to be considerably more stable genetically. Furthermore, the stabilizing effect of interruptions within the repetitive run in long-normal alleles is probably a result of their destabilizing effect on these unusual DNA structures. This has led to the idea that a misalignment between the two repetitive strands during DNA replication which can further stabilized by unusual conformations of repetitive slip-outs, is the basis of repeat instability [5].

3. PROPOSED WORK

The proposed system basically aims at analyzing the DNA structure of the host in comparison with that of the causative organism. It basically consist of 4 modules. The basic architecture is as shown in Figure 1.

There are basically four modules. They are:

- Data Validity Checking
- Data Sequencing
- Pattern Matching
- Disease Detection

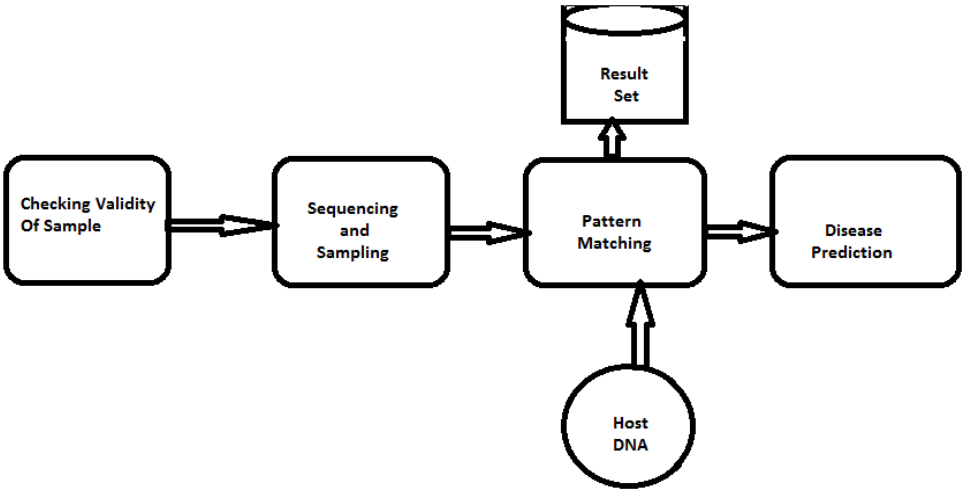

Figure 1: Basic Architecture

Initially the validity of data sample is checked. This checks for the presence of valid bases within the obtained sample. This can identify if there is any invalid mutations present or not within the new sample introduced. The pattern recognizing mechanism makes use of sampling the whole sequence and then generating the corresponding hash key pair.

Once the validity is checked then comes the sequencing and sampling of the DNA data of the sample. This is done using the basic DNA Codon. The genetic code is traditionally represented as an RNA codon table. This is because when proteins are made in a cell by ribosomes, it is mRNA that directs protein synthesis. The mRNA sequence is genetically determined by the sequence of genomic DNA. With the latest enhancements in computational biology and genomics, most genes are now discovered at the DNA level. This has led to increasing use of DNA codon table. The DNA codons in such tables occur on the sense that DNA strands and are arranged in a $5' \rightarrow 3'$ direction. This codon table is universal for all organisms.

The next involves processing the incoming (disease causing) DNA there by analyzing its pattern and mutation percentage. Also the novel algorithm proposes a way to match the host DNA with that of the incoming one. This is done by evaluating the structural similarity, sequence similarity as well as pattern match of both the DNA sequences. This thus checks for the possibility of DNA replacements between the 2 organisms. This is actually a measure of whether the host will get affected by the particular incoming organism or not. It is important to calculate both the structural as well as sequence similarity because DNA as well as RNA being biomolecules the similarity needs to be checked in both sequence as well as structure. The similarity in one measure does not ensure that the other one will also be similar. For replacements to take place we need to ensure both structural as well as sequence similarity. To be more precise a concept of average similarity can be introduced in case of biomolecules like DNA and RNA combining both sequence similarity as well as structural similarity.

For setting up a desired match level between the host and causative organism, different flag parameters are considered. These flag parameters are considered based on the basic DNA properties. DNA has a number of special physical and chemical properties which are important to its structure and functioning. In living organisms such as humans, DNA exists as a pair of molecules ie in a double helix fashion rather than a single molecule. This is kept stable by hydrogen bonds, which can be found between the bases attached to the two strands. A long polymer like DNA is made up of smaller units called nucleotides. Each nucleotide in turn consists of a phosphate group, a sugar and a nitrogenous base. To get a sense of just how long a DNA polymer is, just consider that one human chromosome is several hundred million base pairs long. The four basic parameters considered here are:

A. *Total Strand Length:* The total length of sequence of both host and causative organism is considered to get an idea about accuracy of the total genetic element considered.

B. *DNA Structure and Symmetry:* DNA is made up of molecules called nucleotides. Each nucleotide contains a phosphate group, a sugar group and a nitrogen base. The four types of nitrogen bases are adenine (A), thymine (T), guanine (G) and cytosine (C). The order of these bases is what determines DNA's instructions, or genetic code. This is calculated by counting each elements on the DNA structure.

C. *Base Pairing:* Base pairing basically defines the copying mechanism of DNA was something that dragged great attention when it was discovered. In DNA, bases are specific in that an adenine base, for example, only pairs with a thymine base. Following on that premise, a cytosine base will only bond to a guanine base. This base pairing is known by the name complementary base pairing. The concept is quite simplistic but it is significant for DNA. Here the match pair is calculated by getting (traversing) through the 2 DNA parts and finding all the match elements and representing it either as number or match percentage.

D. *DNA Grooves:* DNA has two kinds of grooves that play important roles in its functioning. Major and minor grooves are structures to allow for necessary proteins in your body to make contact with bases. They are also significant for cell development and communication. As such, these DNA grooves seen in the structure of DNA facilitate the binding of proteins like transcription factors, which then serves to keep the cellular processes occurring effectively within your body.

Based on evaluation of these four base parameters a match level is set up between the two organismic groups. Then based on this we can relate the possibility of how this organism may affect different parts of body.

The next step involves the prediction of disease that the host will be getting in near future. This involves the modification of the algorithm to the next level that makes it predict the disease. Further a different causative organism can come in future which has just few variations in the genetic composition as compared to the current organism. This few gene variations can be predicted using genetic algorithm. Genetic Algorithms (GAs) are adaptive heuristic search algorithm based on the evolutionary ideas of natural selection and genetics. As such they represent an intelligent exploitation of a random search used to solve optimization problems. Although randomized, GAs are by no means random, instead they exploit historical information to direct the search into the region of better performance within the search space. The basic techniques of the GAs are designed to simulate processes in natural systems necessary for evolution, especially those follow the principles first laid down by Charles Darwin of "survival of the fittest". The basic steps involved in genetic algorithm are as given below.

- Initially a random population of chromosomes is generated. Here the genetic structure of causative organism is chosen as a chromosome.
- Then the fitness function is evaluated for a chromosome. Here we check whether the corresponding organism can exist in current environmental conditions.
- Then new population is created by the process of selection, crossover as well as mutations.
- Use the new generated population for a further run of algorithm if necessary.
- When end condition is satisfied return the best solution so far identified.

The selection process selects parent chromosome from a population based on fitness. Crossover is done to create new offspring. If no crossover is performed offspring will be exact copy of parents. Mutation is done to introduce new variants into the population. This can bring both good as well as harmful variants. Genetic

algorithm is better than conventional AI in that it is more robust. Unlike older AI systems, they do not break easily even if the inputs changed slightly, or in the presence of reasonable noise. Also, in searching a large state-space, multi-modal state-space, or n-dimensional surface, a genetic algorithm may offer significant benefits over more typical search of optimization techniques.

This novel method can give the disease prediction more accuracy. Also the mechanism will show the percentage of correction that the visitor DNA had done to the host DNA. This can be used by a medical examiner to determine the mode of treatment as well as the design of drugs for a new disease that comes into the arena.

4. CONCLUSION

Through this work a new prediction mechanism is introduced which actually does a DNA analyzing for predicting the occurrence of a disease. This also identifies the percentage of correction in the genetic composition that can lead to occurrence of a disease. This basically used the concept of genetic material replacements for the prediction mechanism. The proposed method will help to predict the upcoming disease there by helping us to create vaccines, this method will be more correct if both the DNA and RNA of the visitor and host will be provided. Hence it will be more effective if both are provided. Further if the obtained match is linked with different vital organs such that the percentage that they may get affected can be found out. So this inference can help a medical examiner to take more effective treatments.

REFERENCES

[1] Barbara E. Stranger and Emmanouil T. Dermitzakis " From DNA to RNA to disease and back: The 'central dogma' of regulatory disease variation", Human Genomics, 2006.

[2] Y. Lee, M. Kim, J. Han, K.H. Yeom, S. Lee, S.H. Baek, And V.N. Kim, Microrna Genes Are Transcribed By RNA Polymerase II, The EMBO Journal (2004) 23, 4051-4060.

[3] Dimitrios Kleftogiannis "YamiPred: A novel evolutionary method for predicting pre-miRNAs and selecting relevant features", IEEE Transactions on Computational Biology and BioInformatics, 2015.

[4] B. Sanghamitra And M. Ramkrishna, Targetminer: Microrna Target Prediction With Systematic Identification Of Tissue-Specific Negative Examples, Bioinformatics, Vol. 25 No. 20 2009.

[5] Mirkin SM, Expandable DNA repeats and human disease, Nature 2007 Jun 21;447(7147):932-40.