



International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

www.iasir.net

Using Phylogenetic Trees for Disease Diagnosis

Shamsudduha Tabish¹ and Prof. Satish Kumbhar²

Department of IT & Computer
College Of Engineering,
Shivaji Nagar, Ganesh Khind,
Pune, Maharashtra, India

Abstract: The Phylogenetic Tree is a tool for tracking the process of evolution by looking into the changes in the genome sequences under study. This tree is a graphical representation of the evolutionary relationships among multiple genes or organisms. We apply the same principle of phylogeny to find out what disease an individual is suffering from. In our method the multiple sequence alignment is applied to a set of omic (Genomic/Proteomic) sequences of {the patient, the family members of the patient and the diseased sequences} and then the similarity in the omic sequences of patient's family members is found along with the loci of each common nucleotide/amino acid, and the dissimilar nucleotides/amino acid at respective loci are discarded also from the patient's & diseased sequences. Finally we create a phylogenetic tree from these sequences which can now be used to visualize the distance among the patient's genome sequence and the diseased genome sequences.

Keywords: Disease diagnosis, evolution, medical diagnosis, Phylograms, cladograms, Phylogenetic trees, Multiple Sequence Alignment

I. Introduction

Genome sequence is the combination of various genes found in an organism. These genes comprise of nitrogen bases (chemical structures) namely Adenine, Thymine, Guanine and Cytosine represented by 'A', 'T', 'G', and 'C' characters respectively. The gene sequences are known to be responsible for complete coding an organism. All the differences among the organisms are because of the difference among these sequences. The human genome is 99.9 percent similar among all the human beings, only 0.1 percent is responsible for all the differences among us ^[13]. The genome is known to be responsible for the appearance and functioning of the human body. This genome is unique to every individual that is why all of us differ from each other in some way or the other.

A. Single Nucleotide Polymorphism:

Single Nucleotide Polymorphism also known as SNP (Snip) is a change of single nucleotide in the genome a particular locus. If such a variation at a single locus is found common in more than 1% of the population, only then it is considered as SNP. Around 90% of the variation in the genome is because of SNPs ^[2]. SNPs are scattered across the human genome by an approximate average of one SNP per thousand base pairs, these SNPs directly affect the gene product that is the protein ^[7]. Sequence variations in the genomes exist at defined positions and are responsible for phenotypic characteristics, including a person's tendency towards complex diseases like heart disease and cancer ^[9].

B. Multiple Sequence Alignment:

Multiple Sequence Alignment (MSA) is usually the alignment of three or more nitrogen base sequences or Nucleic acid sequences of similar length. Homology can be inferred from the output and the evolutionary relationships between the sequences studied. Usually protein sequences are aligned using multiple sequence alignment to find out the relationship among them. The multiple sequence alignment tools compare these sequences and try to correlate each other by introducing gaps in the sequences in order to match these sequences.

C. Phylogenetic Trees:

The evolutionary tree or cladograms were traditionally used to draw evolutionary relationship among the organism; a more modern version of the same is "phylo-genetic tree" which uses gene / protein sequences to draw the evolutionary relationship. These trees dictate the relationship among the organisms based on the similarity and dissimilarity among the nucleotide or nucleic acid sequences.

The tree construction can be done through variety of tree-building methods which include methods based on distances, likelihood and characters. After a phylogenetic tree is constructed, it is important to test its accuracy which refers to the degree to which a tree is close to the true tree.

II. Our Method

Although the diagnosis is possible with gene sequences, protein sequences, and the RNA sequences, but for this paper we will stick to gene sequences. Our method is based on the similarity that the human beings are having in their gene sequences and the assumption that any change in the gene sequence at the loci where the nitrogen bases are usually common in all the human beings is responsible for the abnormality an individual is having.

Inputs Required: Patient's gene sequence, A few of Patient's family member's gene sequences, diseased gene sequences.

We consider patient's family member's sequences for analysis since their gene sequences are most close to the patient's gene sequence, with the help of these sequences we try to find out which mutation in the sequence of the patient is responsible for the disorder. To diagnose the disease we need to compare the sequence of the patient with the gene sequences of diseased genomes. To reduce the time required for diagnosis (through computer processing) we suggest to find out the probable diseases the patient might be suffering from based on the symptoms.

In our method we find out the common nucleotides among the patient and the family member's gene sequences and discard the dissimilar nucleotides to retain the common nucleotides with respect to their loci. Following are the steps that we suggest to diagnose the disease.

Step 1: Align the gene sequences of

The patient,

The family members of the patient,

And the diseased sequences.

Step 2: Find out the common nucleotides among the family members of the patient, and discard the dissimilar nucleotides from all the sequences (of the patient, patient's family members, and the diseased sequences) from the respective loci after alignment.

Step 3: Now Discard the Patient's family member's gene sequences.

Step 4: Create a phylogenetic tree (we prefer maximum parsimony based phylogenetic tree) based on the sequences we got in the previous step (Modified gene sequences of the patient and the diseased sequences).

Step 5: From this tree we can say that the patient is suffering from a disease which is having least distance from the patient's gene sequence.

III. Example

Let's consider the following hypothetical sequences.

```
>P      TAAAAGTGCCGAAGGAAAA
>F1     TCGAAGTGAAGAGAGAAAA
>F2     TCACAGTGAAGAGAGAAGA
>F3     TCACAGTGAAGAGCGAAAA
>F4     TCAAAGTGAAGAGGGACAA
>D1     TATAAGTGAAGAGGGAAAA
>D2     TCAGCGTGCCGAGGTCGAA
>D3     TCACAGTGAAGACTAAA
>D4     TGAAGTGAAGAGGCTTAA
```

Where P is the Patient, F1, F2, F3 and F4 are close relatives of the patient and D1, D2, D3 and D4 are People suffering from different diseases (Reference sequences).

Now we do apply multiple sequence alignment on these sequences and get the following output.

```
>P      TAAAAGTGCCGAAGGAAAA 19
>D1     TATAAGTGAAGAGGGAAAA 19
>F1     TCGAAGTGAAGAGAGAAAA 19
>F4     TCAAAGTGAAGAGGGACAA 19
>F2     TCACAGTGAAGAGAGAAGA 19
>F3     TCACAGTGAAGAGCGAAAA 19
>D2     TCAGCGTGCCGAGGTCGAA 19
>D3     TCACAGTGAAGA--CTAAA 17
>D4     TGAAGTGAAGAGGCTTAA 19
          *      ***      **      *
```

From the above result we discard the dissimilar nucleotides/characters from the family members sequences & discard the nucleotides/characters at respective loci from the other sequences as follows.

```

>P      TAAAGTGCCGAAGGAAA 19
>D1     TATAAGTGAAGAGGGAAA 19
>D2     TCAGCGTGCCGAGGTCGA 19
>D3     TCACAGTGAAGA-CTAAA 17
>D4     TGAAGTGAAGAGGCTTAA 19
>F1     TCGAAGTGAAGAGAGAAA 19
>F4     TCAGAGTGAAGAGGGACAA 19
>F2     TCACAGTGAAGAGAGAAGA 19
>F3     TCACAGTGAAGAGCGBAAA 19
          ** ***** ** *
  
```

And we get

```

>P      TAAGTGCCGAAGAA
>D1     TAAGTGAAGAGGAA
>D2     TCCGTGCCGAGTCA
>D3     TCAGTGAAGA-CTA
>D4     TGAGTGAAGAGCTA
>F1     TCAGTGAAGAGGAA
>F4     TCAGTGAAGAGGAA
>F2     TCAGTGAAGAGGAA
>F3     TCAGTGAAGAGGAA
          *****
  
```

Further we ignore the close relatives' sequences and construct a phylogenetic tree based on the rest of the sequences.

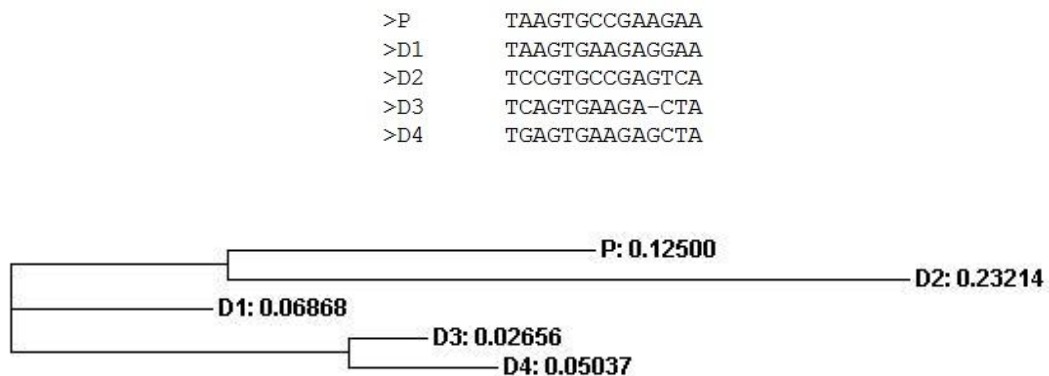


Fig1. Phylogenetic tree depicting relationship among the patient's gene sequence and different diseased sequences

The tree shown in the Fig1. Depicts that the patient is suffering from the disease D2.

IV. Conclusion

The phylogenetic trees can be utilized to diagnose the disease after multiple sequence analysis on the various sequences. This will improve the diagnosis process and hence the process of treatment can be accelerated.

V. Future Work

We need to test this algorithm on different data sets from various sources such as dbSnp, HapMap, 1000 Genome project. We will improve this algorithm to target the non-genetic disease.

VI. References

- [1] Mones Abu-Asab, Mohamed Chaouchi, and Hakima Amri. "Evolutionary medicine: A meaningful connection between omics, disease, and treatment". *Proteomics Clin Appl*. Author manuscript PMC 2008 May 5.
- [2] Anthony J. Brookes "The essence of SNPs" *Gene* 234 (1999) 177-186

- [3] The International SNP Map Working Group, "A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms", *Nature* | Vol. 409 | 15 February 2001
- [4] Alan R. Templeton, "A Cladistic Analysis of Phenotypic Associations with Haplotypes Inferred from Restriction Endonuclease Mapping or DNA Sequencing. V. Analysis of Case/Control Sampling Designs: Alzheimer's Disease and the Apoprotein E Locus", *Genetics Society of America*, February 4, 1995
- [5] Samuel A Smits, Cleber C Ouverney, "Phylometrics: a pipeline for inferring phylogenetic trees from a sequence relationship network perspective", *Smits and Ouverney BMC Bioinformatics* 2010, 11(Suppl 6):S18
- [6] Walter Salzburger, Greg B. Ewing And Arndt Von Haeseler "The Performance Of phylogenetic algorithms in estimating haplotype genealogies with migration" *Molecular Ecology* (2011) 20, 1952–1963
- [7] AfshinAhmadian, BabackGharizadeh, AnnaC.Gustafsson, FredrikSterky, PålNyre'n, MathiasUhle'n, and Joakim Lundeberg, "Single-Nucleotide Polymorphism Analysis by Pyrosequencing", *Analytical Biochemistry* 280, 103-110(2000) doi: 10.1006/abio.2000.4493
- [9] Kitts A, Sherry S. The Single Nucleotide Polymorphism Database (dbSNP) of Nucleotide Sequence Variation. 2002 Oct 9 [Updated 2011 Feb 2]. In: McEntyre J, Ostell J, editors. *The NCBI Handbook* [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2002-. Chapter 5. Available from: <http://www.ncbi.nlm.nih.gov/books/NBK21088/>
- [10] Bryan Bergeron "Bioinformatics Computing", November 19, 2002, ISBN: 0-13-100825-0
- [11] Alexander Isaev "Introduction to Mathematical Methods in Bioinformatics", Springer-Verlag Berlin Heidelberg 2006
- [12] Arthur M. Lesk "Introduction to Bioinformatics", Oxford University Press, 2002, ISBN (Pbk) 0 19 925196 7
- [13] Alison Van Eenennaam "Genomics and Animal Agriculture" April 2008
http://animalscience.ucdavis.edu/animalbiotech/Outreach/Genomics_and_animal