International Journal of Swarm Intelligence and Evolutionary Computation

Shukla and Chaubey, Int J Swarm Intel Evol Comput 2017, 6:2
DOI: 10.4172/2090-4908.1000163

Research Article OMICS International

Analysing the Genetic Diversity of Commonly Occurring Diseases

Rati Shukla[1]* and Punit Kumar Chaubey[2]

[1]GIS CELL, Motilal Nehru National Institute of Technology, India
[2]Department of Computer Science and Engineering, Motilal Nehru National Institute of Technology, India

Abstract

It is commonly alleged that the existence of all organisms present on this earth have their point of convergence in a common gene pool. The current species passed through an evolutionary process which is still underway. The theoretical assumptions relating to the common descent of all organisms are based on four simple facts: First, they had wide geographic dispersal, second, the different life forms were not remarkably unique and did not possess mutually exclusive characteristics, third, some of their attributes which apparently served no purpose had an uncanny similarity with some of their lost functional traits and last, based on their common attributes these organisms can be put together into a well-defined, hierarchical and coherent group, like a family tree. Phylogenetic networks are the main tools that can be used to represent biological relationship between different species. Biologists, Mathematicians, Statisticians, Computer Scientists and others have designed various models for the reconstruction of evolutionary networks and developed numerous algorithms for efficient predictions and analysis. Even though these problems have been studied for a very long time, but the computational model built to solve the biological problems fail to give accurate results while working on real biological data, which could be due to the premises on which the model is based. The objective of this paper is to test and analyse the transmission of commonly occurring diseases to fit into more realistic models. The problems are not only important because we need to know how they came into existence and how they migrated, but also helpful for the treatment of such diseases and drug discovery.

Keywords: Phylogenetic reconstruction; Algorithms; Genetic diversity; Substitution model; Biological relationships

Introduction

Phylogeny is the evolutionary history of a genetically related species or a group of life forms. In excellent contingency this may really be distinguished, the most significant point of biological process reconstruction is to explain biological relationships that species or group descend from common ancestry. Thus we can easily say that phylogenetic analysis plays a significant role in modern biological applications. These applications are ancestral sequence reconstruction, multiple sequence alignment, recombination and hybridization [1,2]. In phylogenetic, relationships are represented as a branching diagram known as tree. These branches are known for lineage distance which may split into independent branches or hybridize or sometimes may even become extinct. There are three divisions of relationship-Monophyly, Paraphyly and Polyphyly [3]. Monophyly and Paraphyly groups have a single ancestral origin. Monophyletic groups always have all the descendants from the same origin. On the other hand, if one lineage emerging from a monophyletic group is removed, then the group is named as para-phyletic. In polyphyletic groups characters absent from the most recent ancestor lead to the formation of this group type [4]. Each homologous sequence could be treated as a single character trait. This single character trait is responsible for the phylogenetic reconstruction. Change of these character traits is also helpful in phylogenetic inference. For this to work, one needs to arrange DNA sequences. This process is called sequence alignment. So, it is believed that phylogenetic reconstruction among dataset materials may be a useful aid in the understanding of the existence of a disease, disease migration pattern, treatment of such diseases and drug discovery. An assortment of distinctive methodologies is currently available for investigation of genetic diversity. These methods depend on performance data, pedigree data, agronomic data, biochemical data, morphological data, and DNA-based data. Since the objective of this study is to test and analyse the transmission of commonly occurring diseases to know how they came into existence and how they migrated to find the treatment of such diseases. In this study to get the moderately correct and unbiased estimates of phylogeny we intend to focus on the following:

(i) Sampling strategies of the datasets

(ii) Genetic distance

(iii) Genetic relationships

(iv) Reconstruction of phylogeny and

(v) Bootstrapping

Motivation and background of this research

Problem statement, problem description and problem formulation: Topical developments in genomics research have furthered progress in the discovery of vulnerability genes and fuelled potentials about opportunities of genetic profiling for personalizing medicine. The intricacy of complex diseases may eventually limit the prospects for precise prediction of disease in asymptomatic individuals as unscrambling their comprehensive causal pathways may be impossible. One of the architypes in complex genetics is that if we are able to identify genetic variants with resilient effects, either on their own or in interaction with other variants or with environmental factors, i.e., gene–gene or gene–environment interaction, the genetic prediction of common diseases can be significantly improved if we are able to identify genetic variants with. If we are capable to understand the vital genetic and environmental factors in pivotal mechanisms of the disease,

***Corresponding author:** Rati Shukla, GIS CELL, Motilal Nehru National Institute of Technology, Allahabad, India, Tel: +919455604876; E-mail: mca.rati@gmail.com

Received August 08, 2016; **Accepted** July 11, 2017; **Published** July 18, 2017

Citation: Shukla R, Chaubey PK (2017) Analysing the Genetic Diversity of Commonly Occurring Diseases. Int J Swarm Intel Evol Comput 6: 163. doi: 10.4172/2090-4908.1000163

perfect prediction of disease may be achieved. A minimally sufficient set of conditions and events leading to disease inexorably was defined. Hence randomness does not exist in Rothman and Greenland models of complete causal mechanisms. CAG extensions in the huntingtin gene are a comprehensive and adequate cause for the growth of the disease regardless of the fact that there may be genes modifying age of onset. For common diseases resulting from various genetic and environmental causes, the complete causal mechanisms are by far more intricate.

They will not only consist of a huge number of various component causes, but a particular disease may also grow from various causal mechanisms.

For example, four different risk variants in diverse genes, may lead to a complex disease, but even in the absence of one of the risk variants the disease may inexorably occur in the presence of four other risk variants with an environmental risk factor. Hence for complex diseases there are various discrete risk factors leading to disease growth, along with major single risk factors incipient in multiple combinations.

Ascertaining comprehensive causal mechanisms of common diseases entails the identification of precise combinations of causal factors among all possible combinations, viz. identifying those combinations that inevitably lead to disease. Since number of multi factorial diseases cause because of a complex interplay of various genetic and non-genetic factors, the number of latent combinations of these many factors is enormously large and easily outnumbers even the size of large cohorts or consortia.

This fact has two implications. Firstly, it's extremely grim to prove that the profiles that are found only among cases essentially are complete causal mechanisms, as it is tremendously unlikely that the same amalgamation of risk factors will be found in more than one person. Second, even if precise combinations could be acknowledged as complete cause mechanisms, then still its expediency for the prediction of common disease is inadequate.

When amalgamations of risk factors are 'unique', only a few other persons in the world may have exactly same profile.

The uniqueness of profiles is not astonishing in the field of genetics, as it is the basic foundation for contemporary practice in forensic genetics and paternity testing. Based on a standard set of 13 specific short tandem repeat regions, forensic analysts of the Federal Bureau of Investigation (FBI), created DNA profiles (fingerprints). The odds for two individuals having accurately the same profile are less than 1 in a billion.

Potential suspects may be identified when their 13-loci DNA profile matches the evidence left at the crime scene. If this concept of individuality is true for forensic and legal applications, it will also hold in medicine. One obvious exemption is the inheritance of genetic profiles within families, but also in this case the probability of sharing the same combination of multiple risk genotypes is likely too minor to be useful for disease prediction.

The inadequate value of prophecy of future occurrences based on a precise combination of numerous causal factors is not elite to medicine, but largely encountered in the prediction of complex events, such as the prevention of catastrophes and disasters.

This fact unquestionably donated to the capsizing, but does not fully explain the catastrophe as several earlier efficacious crossings with open bow doors have been reported. The value for the prediction of future tipping is virtually zero even when the cause of the disaster is known completely, as particular combination is supposed to be rare and a number of other factors may contribute. Even faultless considerate of causal pathways may not result in candid prediction of intricate diseases as is possible for Huntington Disease. Prognostic testing of common diseases, either based on genetic variants only or in amalgamation with environmental risk factors, will remain based on statistical models.

Methods

The conduction of infectious diseases is an inherently ecological process. It involves interactions among different species. Although current studies have been able to shed light on the diversity of disease origin, the mechanisms underlying these effects stay unclear in several cases. In this work we conducted examination on a worldwide scale to analyse whether the diversity of human diseases, some of them in charge of high rates of morbidity and mortality. In this study we also guaranteed to get the list of highly conserved motifs. The program is sure to report all sets of motifs with lowest parsimony scores. These are calculated with regard to the phylogenetic tree relating the different input species. Further the computation of genetic diversity within the subpopulation and entire population easily show the presence of disease and its variances. At last reconstruction of phylogenetic network helps in tracing phylogeny of the diseases. Subsequent to controlling for direct puzzling impacts applied by distinctive strands of information taken, our discoveries demonstrate that human disease increments occur with the assorted diversity and structure of disease types.

The use of statistical tools and techniques, such as, computing the selection of substitution model, computing mean diversity within the subpopulation as well within entire population is an important requirement; followed by reconstruction of phylogenetic network using any efficient algorithm such as ReTF [5]. After reconstruction process bootstrapping is performed to understand the results derived from different types of data sets. It focuses on statistical analysis techniques with the already defined algorithm. It is more helpful with both features in analysis of genetic diversity at the intraspecific level in disease finding.

In this study commonly occurring virus datasets from National Centre for Biotechnology Information have been used to perform the analysis. Many viruses that were earlier present in only a few parts of the world are now spreading throughout the world; these include HIV1, Ross River virus, H1N1, Rift Valley Fever virus, Ebola, Zika virus, Japanese encephalitis virus, louping ill virus, chickungunya virus and West Nile virus. Spread of these viruses is related to climate change and other environmental factors. In this study our emphasis will be on HIV1, H1N1 and Ebola viruses. Viruses are thought-out to be polyphyletic; it means that can they have many evolutionary origins. Keeping this in mind different strands of datasets are taken for analysis. Once a dataset is ready we run multiple sequence alignments. There are widely available software's which are used for alignments e.g. ABA, ClustalW, DNA Alignment, MUSCLE, Phylo, T-Coffee. In this study we have used ClustalW to arrange the sequences. MSA (Multiple sequence alignments) are used to arrange three or more biological sequences. Initially, when we get the data each strand is of a different length, also at the same time some gaps are present. To align such sequences is a tedious process. Therefore, computational algorithms are used to align such sequences. Once alignment is done, we will be using ReCons algorithm to find the list of most conserved motifs [5]. These conserved motifs depend on p-value calculations after using

PMS5 [6]. The list of motifs with predefined threshold value is used to find the best substitution model. This substitution model describes the process from which a sequences of characters is recast into another set of traits. Substitution model is used for various aims. Primarily it is used to construct the evolutionary trees. Sometimes it is used to simulate sequences to test newly designed algorithms. Many different substitution models are used but the very old and renowned model Generalised Time Reversible (GTR) is widely accepted. This model was created by Simon Tavare in 1986. It is the most general, independent, finite-sites, time reversible model. With the help of motifs list that were the output of ReCons algorithm and on the basis of statistical analysis of substitution model, phylogenetic network is reconstructed Bootstrapping is the process of evaluating the relative strength of the newly-constructed phylogenetic network to the original tree [7]. In this each interior branch of tree is compared. This resampling process is repeated several hundred times and each interior branch is assigned a bootstrap value (Figure 1).

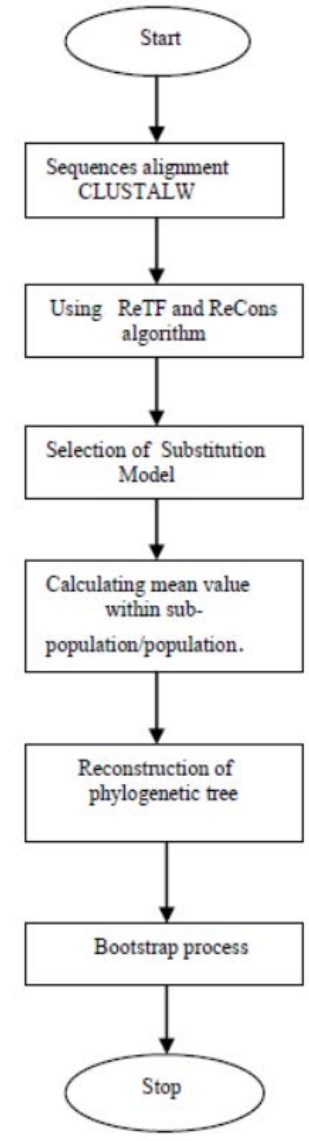

Figure 1: Steps in analysis of viruses' datasets.

Threshold	Databases of Viruses		
	HIV 1 Motifs count	H1N1 Motifs count	Ebola Motifs count
0.993-0.999	11	8	12
0.899-0.992	32	28	26

Table 1: Threshold table for databases.

Results and Discussion

To get correct phylogeny, we consider the following parameters.

Step 1 Analysing the data samples on the basis of motif using (ReCons) algorithm with statistical analysis

DNA sequences are taken from National Centre for Biotechnology Information (NCBI). Sequences are arranged using CLUSTALW algorithm. Once sequences are arranged, PMS5 start sequences alignment CLUSTALW using ReTF and ReCons algorithm selection of substitution model calculating mean value within subpopulation/population. Reconstruction of phylogenetic tree Bootstrap process Stop is used to find the motifs with constrains (l,d) where l is the length of motif and d is the allowed number of mutation. We get hundreds to thousands of motifs. Here ReCons algorithm is used to find the list of most conserved motifs. Predefined threshold is taken into account when looking for the conserved motifs [5] (Table 1).

Step 2 Selection of substitution model

Selection of best substitution depends on Maximum Likelihood fits of 24 different nucleotide substitution models. Models with the most reduced Bayesian Information Criterion (BIC) are acknowledged for representing the best substitution pattern. For every substitution model, Maximum Likelihood value (lnL), Akaike AICc value, information criterion and the parameters are presented in the tables below [8]. Discrete Gamma distribution (+G) having 5 rate division with consideration that a chunk of sites is evolutionary invariable (+I) can be used to present the non-consistency in the evolutionary rates. Gamma shape parameter and invariant sites are depicted at relevant points. There were a total of 1137 positions in HIV, 2300 in H1N1, 4322 in Ebola. These evolutionary analyses were conducted in MEGA6 [9]. An inherent obstacle in evolutionary analysis is the choice of appropriate selection of the best substitution model. AICc or BIC mostly use theoretical metrics. A researcher usually estimates up to three parameters to describe the substitution model [10,11]. The most important factor that should be considered is the rate multiplier which is responsible for the overall substitution model. Secondly, one or more parameters are used to describe the relative rates at which nucleotides replace each other. This is called General Time Reversible (Tables 2-4).

Step 3 Analysing the data samples on the basis of gene diversity within the sub population and within the entire population

Genetic diversity is a necessary feature of all living organisms. It provides the resource for the progressive adaption of the population to ever-changing setting. It describes naturally genetic difference among individuals of the same species. This is also called quasi-species. These variations help in the survival of genes even after climatic changes. Also there is genetic drift which can be described as increase or decrease of population by chance over a period of time. Genetic drift is common issue after population bottleneck. Initially while analysing the genetic diversity, sequencing of DNA clones are obtained from multiple plaques. These plaques were descending from plaque-purified genomes. This approach is further used to calculate mutation rate used

Model	Parameters	BIC	AICs
HKY+G	32	11752.89	11505.2
TN93+G	33	11762.41	11506.98
HKY+G+I	33	11762.64	11507.21
GTR+G	36	11769.78	11491.15
TN93+G+I	34	11772.15	11508.99
GTR+G+I	37	11779.53	11493.16
HKY+I	32	11792.57	11544.88
TN93+I	33	11802.31	11546.88
GTR+I	36	11808.68	11530.05
HKY	31	11814.78	11574.82
TN93	32	11824.32	11576.63
GTR	35	11832.74	11561.84
T92+G	30	11842.88	11610.66
T92+G+I	31	11852.62	11612.67
T92+I	30	11880.42	11648.2
T92	29	11899.82	11675.34
K2+G	29	11958.68	11734.2
K2+G+I	30	11968.42	11736.2
K2+I	29	11996.94	11772.46
K2	28	12026.42	11809.68
JC+G	28	12272.59	12055.85
JC+G+I	29	12282.33	12057.85
JC+I	28	12305.65	12088.91
JC	27	12314.24	12105.24

Table 2: Substitution model analysis for HIV1 virus.

Model	Parameters	BIC	AICs
T92	29	4267.496	4036.756
HKY	31	4271.173	4024.526
T92+I	30	4277.456	4038.762
T92+G	30	4277.456	4038.762
TN93	32	4280.265	4025.665
HKY+G	32	4281.133	4026.532
HKY+I	32	4281.133	4026.532
T92+G+I	31	4287.415	4040.768
TN93+I	33	4290.225	4027.671
TN93+G	33	4290.225	4027.671
JC	27	4290.449	4075.617
HKY+G+I	33	4291.092	4028.538
K2	28	4294.018	4071.232
TN93+G+I	34	4300.184	4029.677
JC+I	28	4300.408	4077.622
JC+G	28	4300.409	4077.623
K2+I	29	4303.977	4073.237
K2+G	29	4303.978	4073.237
GTR	35	4308.689	4030.229
JC+G+I	29	4310.368	4079.628
K2+G+I	30	4313.937	4075.243
GTR+G	36	4318.631	4032.219
GTR+I	36	4318.648	4032.236
GTR+G+I	37	4328.591	4034.226

Table 3: Substitution model analysis for H1N1 virus.

for genetic diversity. Genetic diversity is important for two reasons. If the population of an organism contains a large gene pool, then we can easily analyse its chances of surviving and flourishing. A small gene pool has limited genetic variability. Some people could have acquired characteristics making them especially resistant to diseases. Sometimes

Model	Parameters	BIC	AICs
TN93+I	33	55541.97	55193.84
HKY	31	55542.28	55215.26
TN93	32	55542.5	55204.92
HKY+G	32	55543.08	55205.5
TN93+G	33	55543.32	55195.19
TN93+G+I	34	55554.14	55195.46
HKY+I	32	55554.7	55217.12
HKY+G+I	33	55555.46	55207.34
GTR	35	55577.19	55207.97
GTR+G	36	55578.18	55198.41
GTR+I	36	55585.04	55205.26
GTR+G+I	37	55590.71	55200.39
T92	29	55606.82	55300.89
T92+G	30	55607.73	55291.25
T92+I	30	55618.6	55302.12
T92+G+I	31	55620.26	55293.23
K2	28	56192.22	55896.84
K2+G	29	56192.99	55887.06
K2+I	29	56204.75	55898.82
K2+G+I	30	56205.25	55888.77
JC	27	56479.4	56194.57
JC+G	28	56480.78	56185.4
JC+I	28	56491.93	56196.55
JC+G+I	29	56493.32	56187.39

Table 4: Substitution model analysis for Ebola virus.

	Viruses		
Genetic Diversity	**HIV1**	**H1N1**	**Ebola**
Mean Diversity within subpopulation	218.00	117.00	221.00
Mean Diversity in entire population	312.00	227.00	309.00

Table 5: Genetic diversity within sub-population and within entire population.

there is likelihood that they may have different attributes that increase their chances for endurance. In nature, the "fittest" people survive and go ahead to reproduce-Darwin designated this procedure "natural selection". Secondly, genetic diversity also reduces the incidence of unfavourable inherited traits. When the group is small, the chances to breed within the group increase manifold. This helps to maintain the genetic constitute of the individual (Table 5).

Step 4 Analysing the reconstruction of phylogenetic networks

There are two commonly used approaches for inferring phylogenies. The first approach is phenetic approach. In this approach inference is not drawn regarding any historical relationships, only the distance between species is measured. In order to create tree hierarchal clustering approach is used. Other approach is cladistics, in which all possible paths of evolution are considered. Each node is inferred during the process and choosing an optimal tree according to some model of evolutionary history. In this study we will be focusing on cladistics approach. The algorithm used to reconstruct the phylogenetic network is maximum likelihood.

This method was introduced by Felsenstein [12]. This method doesn't impose any constraint on the constancy of biological process rate among lineages. It assigns quantitative possibilities to mutational events, instead of simply counting them [13]. This methodology compares doable biological process trees on the premise of their ability to predict the discovered information. The tree having the maximum probability of deriving the detected sequences can be picked. To check the correctness of reconstructed tree bootstrap process is used [1].

Phylogenetic Reconstruction Analysis	
Statistical Analysis Method	Maximum Likelihood
Phylogeny Test	
Test of Phylogeny	Bootstrap phylogeny test
No. of Bootstrap	500
Substitution Model	
Substitution Type	Nucleotide
Model	GTR
Performance	
No. of Threads	1

Table 6: Stages of analysis during reconstruction of phylogenetic networks.

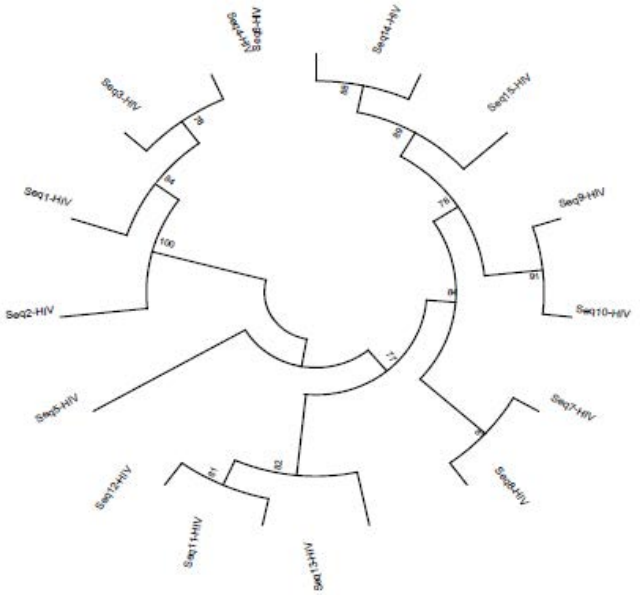

Figure 2: Bootstrap analysis of HIN1 using ReTF algorithm.

Bootstrap is basically the confidence level of phylogenetic network. In this process sampling is done for each n nucleotides dataset. From each succession dataset, n nucleotides are haphazardly picked with replacements, giving rise to a row of b columns each. From this new matrix, a tree is reproduced. After that topology of this tree is compared to that of the original tree. Each inside branch is given a score value that represents confidence of reconstruction of phylogeny. If the bootstrap value for a given interior branch is 65% or higher, then the topology at that branch is considered "correct". Following analysis is performed on the HIV1, H1N1 and Ebola datasets (Table 6).

The results of reconstruction of phylogenetic network using ReTF algorithm is shown in Figures 2-4 [5].

Conclusion

The conduction of infectious diseases is an inherently ecological process. It involves interactions among different species. Although current studies have been able to shed light on the diversity of disease origin, the mechanisms underlying these effects stay unclear in several cases. In this we conducted examination on a worldwide scale to analyse whether the diversity of human diseases, some of them in charge of high rates of morbidity and mortality. In this study we also guaranteed to get the list of highly conserved motifs. The program is sure to report all sets of motifs with lowest parsimony scores. These are calculated

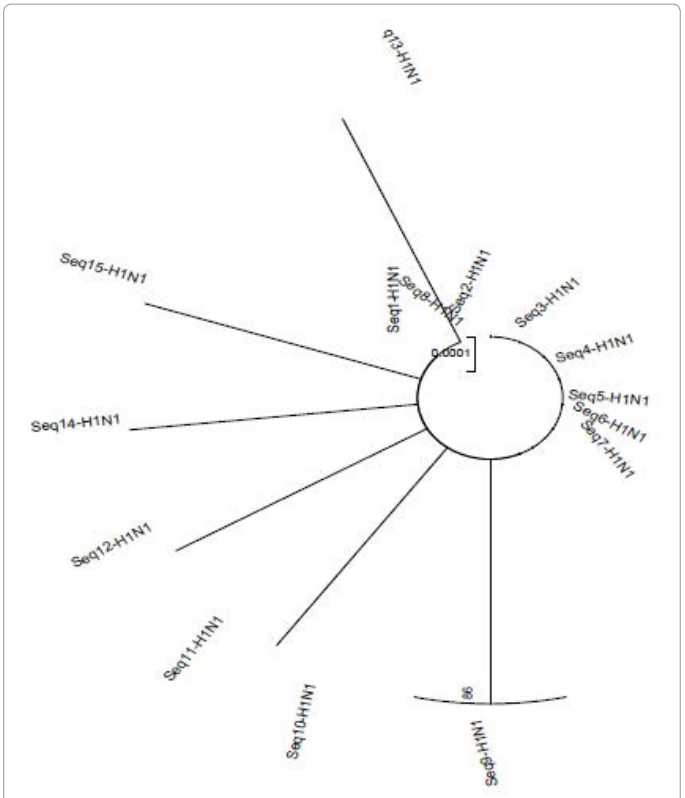

Figure 3: Bootstrap analysis of H1N1 using ReTF algorithm.

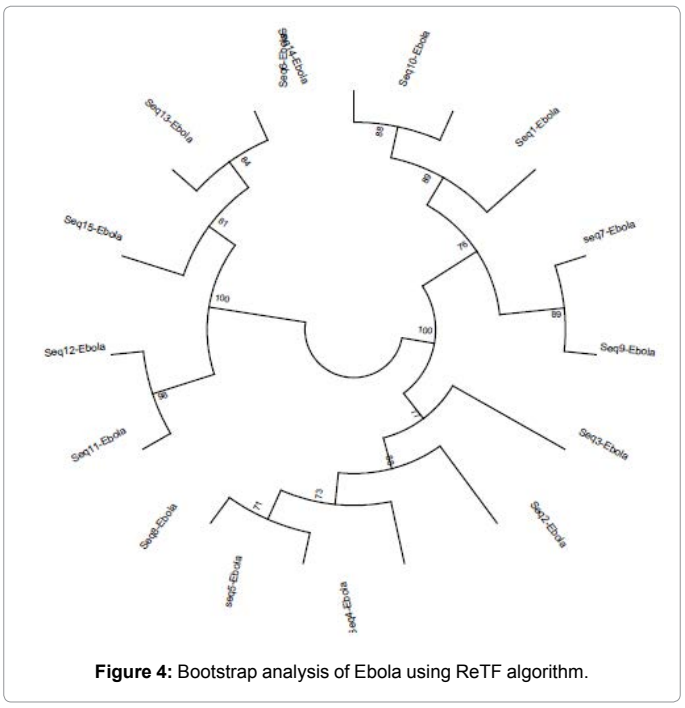

Figure 4: Bootstrap analysis of Ebola using ReTF algorithm.

with regard to the phylogenetic tree relating the different input species. Further the computation of genetic diversity within the subpopulation and entire population easily show the presence of disease and its variances. At last reconstruction of phylogenetic network helps in tracing phylogeny of the diseases. Subsequent to controlling for direct

puzzling impacts applied by distinctive strands of information taken, our discoveries demonstrate that human disease increments occur with the assorted diversity and structure of disease types.

References

1. Magner LN (2002) A history of the life sciences, revised and expanded. CRC Press.
2. Jain R, Rivera MC, Lake JA (1999) Horizontal gene transfer among genomes: The complexity hypothesis. Proc Natl Acad Sci 7: 3801-3806.
3. Nelson G (1978) Historical biogeography: An alternative formalization. Syst Zool.
4. Schwarz G (1978) Estimating the dimension of a model. Ann Stat.
5. Felsenstein J (1981) Evolutionary trees from DNA sequences: A maximum likelihood approach. J Mol Evol 17: 368-376.
6. Felsenstein J (1985) Confidence limits on phylogenies: An approach using the bootstrap. Evol.
7. Kitching IJ (1998) Cladistics: The theory and practice of parsimony analysis. Oxford University Press.
8. Nei M, Kumar S (2000) Molecular evolution and phylogenetics. Oxford University Press.
9. Dinh H, Rajasekaran S, Kundeti VK (2011) PMS5: An efficient exact algorithm for the (ℓ, d)-motif finding problem. BMC Bioinformatics.
10. Shamita M, Sharma D (2013) Reconstructing phylogenetic network with ReTF algorithm (rearranging transcriptional factor). Bioinformatics and Bioeng (BIBE), 13th International Conference on IEEE.
11. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S (2013) MEGA6: Molecular evolutionary genetics analysis version 6.0. Mol Biol Evol.
12. Shamita M, Sharma D (2014) Detecting history of species using mining of motifs in phylogenetic networks. Proceedings of the 2014 International Conference on Information and Communication Technology for Competitive Strategies.
13. Hurvich CM, Tsai CL (1989) Regression and time series model selection in small samples. Biometrika.

Citation: Shukla R, Chaubey PK (2017) Analysing the Genetic Diversity of Commonly Occurring Diseases. Int J Swarm Intel Evol Comput 6: 163. doi: 10.4172/2090-4908.1000163