CRIMSON PUBLISHERS
Wings to the Research

Open Access
Biostatistics & Bioinformatics

ISSN 2578-0247

Mini Review

Phylogenetic Methods and its Applications

Geetika Munjal[1]*, Sangeet Srivastava[1] and Madasu Hanmandlu[2]

[1]*The North Cap University, India*

[2]*IIT Delhi , India*

***Corresponding author:** Geetika Munjal, The North Cap University, Madasu hanmandlu, Gurugram, Delhi, India, Email: *geetika@ncuindia.edu*

Submission: December 16, 2017; **Published:** February 06, 2018

Abstract

The phylogenetic tree portrays a relationship between sets of species and represents a model of molecular evolution. The current forms of species retain many of their ancestral features, some of which gradually change to help these species adjust to their environment. We provide a review of phylogenetic construction and validation methods and its application in cancer analysis. The need of alignment free methods for sequence comparison is also explored as an alternative to alignment based methods.

Keyword: Phylogenetics; Sequence comparison; Tree validation; Alignment free methods

Abbreviations: DNA: Deoxyribonucleic Acid; RNA: Ribonucleic Acid; MAST: Maximum Agreement Subtree

Introduction

Phylogenetic trees help scientists gain a better understanding of how species have evolved while explaining the similarities and differences amongst them. The phylogenetic study can help in analysing the evolution and similarities amongst diseases and viruses, and further helps in prescribing their vaccines [1]. The methods of phylogenetics are broadly classified as distance based and character based methods [2]. Phylogenetics relies on information extracted from the genetic material such as deoxyribonucleic acid (DNA), ribonucleic acid (RNA) or protein sequences. The species can be expressed as DNA strings which are formed by combining four nucleotides A,T,C and G (adenine, thymine, cytosine and guanine respectively). In literature various string processing algorithms are reported which can quickly analyse these DNA and RNA sequences, and build a phylogeny of sequences or species based on their similarity and dissimilarity [1,3]. A high similarity among two sequences usually implies significant functional or structural likeliness, and these sequences are closely related in a phylogenetic tree.

To get precise information about the extent of similarity to some other sequence stored in a database; we must be able to compare sequences quickly with a set of sequences. For this we need to perform multiple sequence comparison, which can be done using alignment based and alignment free methods. Excellent results can be achieved using alignment based methods when the sequences are closely related and can be aligned reliably, but divergence between the sequences affects the alignment [4]. As the number of datasets in phylogenomics increases, the alignment based methods become unaffordable. Multiple alignments of related sequences may often yield the most helpful information on its phylogeny. However, it can produce incorrect results when applied to more divergent sequence rearrangements [5]. Some computationally intensive multiple alignment methods align sequences strictly based on the order in which they receive them, i.e. the input order, without any considering their relationship. Multiple sequence alignment methods emphasize that more closely related sequences should be aligned first. In cases of sequences being less related to one another, however sharing a common ancestor may be clustered separately [4,5]. This implies that they can be more accurately aligned, but may result in incorrect phylogeny. If the differences among the lengths of sequences are very high, the alignment performance significantly impacts tree generation. Another factor that plays a crucial role in the tree construction is, choice of suitable scoring matrices and gap penalties that apply to a set of sequences used in sequence alignment. Gaps in alignments can be exemplified as mutational changes in sequences including insertions, deletions, or rearrangement of genetic material. For phylogenetic analysis the selected sequences should align with each other along with their entire lengths, or else each should have a common set of patterns or domains which provide a strong indication of evolutionary relatedness. Considering the limitation of alignment based methods alignment free methods are proposed in literature.

Alignment free sequence comparison methods include alternative metrics like k-tuple (k is length of subsequence) and probabilistic methods. In the k-tuple method, a genetic sequence is represented by a frequency vector of fixed length subsequence and the similarity or dissimilarity measures are found based on frequency vector of sub-sequences [6]. The probabilistic methods represent the sequences using the transition matrix of a Markov chain of a pre-specified order and comparison of two sequences is done by finding the distance between two transition matrices [7,8].

The output of sequence comparison is used to generate phylogeny of sequences in form of a tree and to extract information from tree structured data; tree mining is to be performed [9]. Tree mining will help to answer various questions like: whether a tree contains all the information contained in another tree or set of trees, ii) whether the constructed phylogenetic tree is correct, iii) whether the various trees generated contain any common patterns. Several methods already exist in literature to answer these questions some of which are bootstrapping, maximum agreement subtree (MAST), and frequent patterns mining [9]. In bootstrap method new alignment matrix of identical dimensions is created to replicate results of multiple sequence alignment that is called as bootstrap replicate. Each bootstrap replicate is in the form of phylogenetic tree contains same number of species [10]. The branches repeated in maximum bootstrap replicates have high confidence. This method has been very successful among alignment based methods [11]. To find the similarity among trees we can find common patterns within a tree which utilizes the concepts of frequent pattern mining [12]. Some methods are based on finding the minimal number of changes required to transform one tree into another. One of its variation is the maximum agreement subtree [5] that finds the maximum common patterns among binary trees.

Phylogeny is not only limited to find similarity and evolutionary patterns in species, it has been utilized in disease and virus analysis as well [13]. The phylogenetic methods have also been identified to understand cancer progression [14]. The methods of phylogenetics have been widely used for tumour classification as it generates a tumour network and infer their progression pathways [15]. Phylogenetics methods can solve the problem of class prediction by using a classification tree [16]. Phylogenetics is a powerful tool for grouping samples of cancer subtypes, and phylogeny inference algorithms can be used to infer how different cancer subtypes have evolved in the population [17]. With help of deep sequencing phylogenetic methods can help in analysing breast cancer progression. Phylogentic methods as hierarchical clustering give us a deeper understanding of biological heterogeneity among cancer subtype when applied to gene expression data. Some methods of phylogenetic like maximum parsimony makes assumptions about the rate at which characters of sequence data change in different regions of the tree, which may give incorrect result [18].

Conclusion

The review presented three main aspects associated with phylogenetics. The first aspect reviews methods of sequence comparison which are alignment based and alignment free methods. Second part reviews methods of tree mining and tree validation. The third aspect has reviewed application of phylogenetic in cancer analysis. The review has highlighted the challenges associated with sequence comparison methods. Phylogenetic methods can be an important tool in cancer analysis; however the methods chosen play a very critical role cancer analysis and its progression.

Acknowledgement

This study is part of Research project sanctioned under Department of Science and Technology, Delhi, India with grant number: SR/WOS-A/ET-1015/2015(G)

References

1. Lam TTY, Hon CC, Tang JW (2010) Use of phylogenetics in the molecular epidemiology and evolutionary studies of viral infections. Crit Rev Clin Lab Sci 47(1): 5-49.
2. Burr T (2010) Phylogenetic trees in bioinformatics. Curr Bioinform 5(1): 40-52.
3. Moret BME, Warnow T (2002) Reconstructing optimal phylogenetic trees: A challenge in experimental algorithmics. Experimental Algorithmics, LNCS, 2547: 163-180.
4. Vinga S (2014) Editorial: alignment-free methods in computational biology. Brief in Bioinform 15(3): 341-342.
5. Höhl M, Ragan MA (2007) Is multiple-sequence alignment required for accurate inference of phylogeny? Syst Biol 56(2): 206-221.
6. Munjal G, Hanmandlu M, Srivastava S, Gaur D (2017) Assessing and mining of phylogenetic trees. International Journal of Database Theory and Application 10(1) 67-78.
7. Kozarzewski B (2012) A method for nucleotide sequence analysis. CMST 18(1): 5-10.
8. Yang K, Zhang L (2008) Performance comparison between k-tuple distance and four model-based distances in phylogenetic tree reconstruction. Nucleic Acid Res 36(5): 1-9.
9. Zaki M (2005) Efficiently mining frequent trees in a forest: Algorithms and applications. IEEE Transactions on Knowledge and Data Engineering 17(8): 1021-1035.
10. Martin DM, Thatte BD (2013) The maximum agreement subtree problem. Discrete Applied Mathematics 161(13-14): 1805-1817.
11. Chuang HY, Lee E, Liu YT, Lee D, Ideker T (2007) Network-based classification of breast cancer metastasis. Mol Syst Biol 3(140): 1-10.
12. Somarelli J, Ware K, Kostadinov R, Robinson J, Amri H, et al. (2017) PhyloOncology: understanding cancer through phylogenetic analysis. Biochimica et Biophysica Acta (BBA)-Reviews on Cancer 1867(2): 101-108.
13. Desper R, Khan J, Schäffer A (2004) Tumour classification using phylogenetic methods on expression data. Journal of Theoretical Biology 228(4): 477-496.
14. Park Y, Shackney S Schwartz R (2009) Network-based Inference of cancer progression from microarray data. IEEE/ACM Trans Comput Biol Bioinform 6(2): 200-212.
15. Alon N, Chor B, Pardi F, Rapoport A (2010) Approximate maximum parsimony and ancestral maximum likelihood. IEEE/ACM Trans Comput Biol Bioinforma 7(1): 1-7.
16. Sardaraz M, Tahir M, Ikram TA, Bajwa H (2012) Applications and algorithms for inference of huge phylogenetic trees: a review. Am J Bioinforma Res 2(1): 21-26.
17. Kuang C, Liu X, Wang J, Yao Y, Dai Q (2015) Position-specific statistical model of dna sequences and its application for similarity analysis. J Mathemeatical Comput Chem 73: 545-558.
18. Hanmandlu MG, Saini A, Gaur D (2015) Modified k-Tuple method for Construction of Phylogenetic Trees. Trends in Bioinformatics 8(3): 75-85.

For possible submission use the below is the URL Submit Article

How to cite this article: Geetika M, Sangeet S, Madasu H. Phylogenetic Methods and its Applications. Open Acc Biostat Bioinform. 1(1). OABB.000505. 2018. DOI: 10.31031/OABB.2018.01.000505