Phylogenetics Algorithms and Applications

Geetika Munjal, Madasu Hanmandlu and Sangeet Srivastava

Abstract Phylogenetics is a powerful approach in finding evolution of current day species. By studying phylogenetic trees, scientists gain a better understanding of how species have evolved while explaining the similarities and differences among species. The phylogenetic study can help in analysing the evolution and the similarities among diseases and viruses, and further help in prescribing their vaccines against them. This paper explores computational solutions for building phylogeny of species along with highlighting benefits of alignment-free methods of phylogenetics. The paper has also discussed the application of phylogenetic study in disease diagnosis and evolution.

Keywords Phylogenetics · Cancer evolution · Sequence analysis

1 Introduction

Phylogenetics can be considered as one of the best tools for understanding the spread of contagious disease, for example, transmission of the human immunodeficiency virus (HIV) and the origin and subsequent evolution of the severe acute respiratory syndrome (SARS) associated coronavirus (SCoV) [1]. Earlier, morphological traits were used for assessing similarities between species and building phylogenetic trees. Presently, phylogenetics relies on information extracted from genetic material such as deoxyribonucleic acid (DNA), ribonucleic acid (RNA) or protein sequences [2]. Methods used for phylogenetic inference have changed drastically during the past two decades: from alignment-based to alignment-free methods [3]. This paper has

G. Munjal (✉) · S. Srivastava
The NorthCap University, Gurugram, India
e-mail: geetika@ncuindia.edu; munjal.geetika@gmail.com

S. Srivastava
e-mail: sangeetsrivastava@ncuindia.edu

M. Hanmandlu
IIT Delhi, New Delhi, India
e-mail: mhmandlu@gmail.com

Y.-C. Hu et al. (eds.), *Ambient Communications and Computer Systems*,
Advances in Intelligent Systems and Computing 904,
https://doi.org/10.1007/978-981-13-5934-7_17

reviewed various methods under phylogenetic tree construction from character to distance methods and alignment-based to alignment-free methods. A brief review of phylogenetic tree applications is also given in cancer studies.

2 Literature Review

A phylogenetic tree can be unrooted or rooted, implying directions corresponding to evolutionary time, i.e. the species at the leaves of a tree relate to the current day species. The species can be expressed as DNA strings which are formed by combining four nucleotides A, T, C and G (A—adenine, T—thymine, C—cytosine and G—guanine). 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. A high similarity among two sequences usually implies significant functional or structural likeliness, and these sequences are closely related in the phylogenetic tree. To get more 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 the multiple sequence comparison. Dynamic programming concepts facilitate this comparison using alignment methods, but it involves more computation. Moreover, the iterative computational steps limit its utility for long length sequences [3]. Alignment-free methods overcome this limitation as they follow alternative metrics like word frequency or sequence entropy for finding similarity between sequences.

3 Methods of Phylogenetic Tree Construction

Phylogenetic tree generation consists of sequence alignment where the resulting tree reveals how alignment can influence the tree formation. Alignment-based methodologies are probably the most widely used tools in sequence analysis problems [4]. They consist of arranging two sequences: one on the top of another to highlight their common symbols and substrings. An alignment method is based on alignment parameters including insertion, deletions and gaps which play a pivotal role in the construction of the phylogenetic tree. A phylogenetic tree is formed as an outcome of sequence analysis performed on the DNA or RNA strings [5]. Sequence comparison reveals the patterns of shared history between species, helping in the prediction of ancestral states. The comparison of sequences also helps in understanding the biology of living organisms which is required to find similarity and relationship among species. For sequence comparison, we can follow alignment-based or alignment-free methods [3, 6, 7].

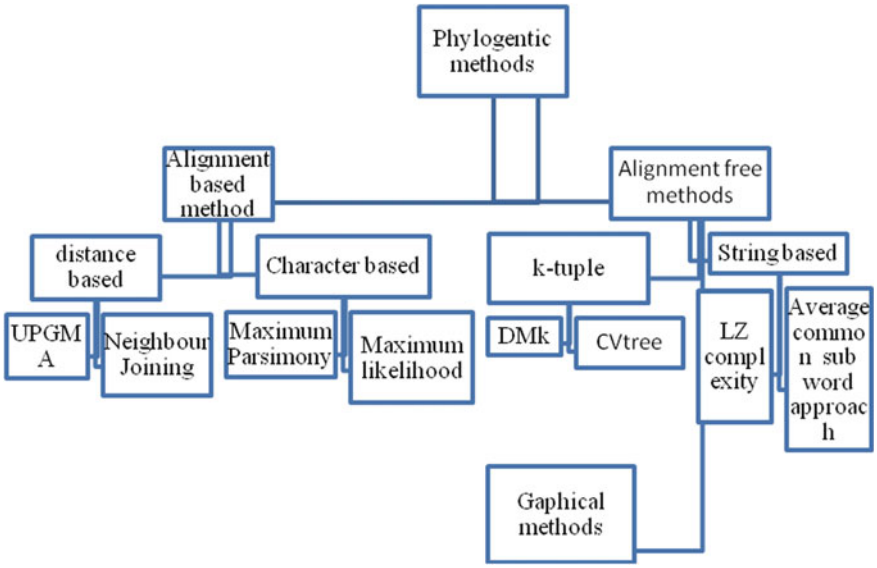

Fig. 1 Hierarchical view of phylogenetic methods[2]

3.1 Sequence Alignment

Sequence alignment is a method to identify homologous sequences. It is categorized as pairwise alignment in which only two sequences are compared at a time whereas in multiple sequence alignment more than two sequences are compared. Alignment-based can be global or local [8, 9]. These alignment-based algorithms can also be used with distance methods to express the similarity between two sequences, reflecting the number of changes in each sequence. Figure 1 gives a hierarchical view of various methods for phylogenetic tree building.

3.2 Character-Based Methods

The character-based methods compare all sequences simultaneously considering one character/site at a time. These are maximum parsimony and maximum likelihood. These methods use probability and consider variation in a set of sequences [10]. Both approaches consider the tree with the best score tree, which requires the smallest number of changes to perform alignment. Maximum parsimony method suffers badly from the long-branch attraction and gives the least information about the branch lengths [10]. In such cases, if two external branches are separated by short internal branches, it leads to the incorrect tree. Some of the salient features of character-based methods are mentioned in Table 1.

Table 1 Comparison of different phylogenetic tree construction methods

Method	Advantage	Disadvantage	Other information
Maximum parsimony	Appropriate for very similar sequences and a small number of sequences	Very time-consuming as it tests all possible trees Parsimony may fail for diverged sequences Suffers from the long-branch attraction	Predict the evolutionary tree that minimizes the number of steps required to generate the observed variation in the sequences It is built with the fewest changes required to explain (tree) the differences observed in the data
Maximum likelihood	Suitable for very dissimilar sequences We can formulate hypothesis about evolutionary relationships More accurate phylogenetic trees can be constructed for a small number of taxa in a reasonable time frame	A slow search algorithm will lead to slow response Takes a long time for large datasets	It tries to find a model that has the highest probability to generate the input sequence under a given evolutionary model
Neighbour joining	Faster than the character-based method They are fast and can be used with a variety of models	Conversion from sequence data to distance data leads to loss of information	Provides an unrooted tree and a single resultant tree
UPGMA	Reliable for related sequences	Evolution rate is constant in all branches	UPGMA provides rooted tree
Fitch Mangrolish	Less sensitive to variations in evolutionary rate	Dependent on the model used to obtain the distance matrix	

3.3 Distance-Based Methods

Distance-based methods use the dissimilarity (the distance) between the two sequences to construct trees. They are much less computationally intensive than the character based methods are mostly accurate as they take mutations into count. For tree generation, generally, hierarchical clustering is used in which dendrograms (clusters) are created. Table 1 briefly compares various phylogenetic tree construction methods.

4 Alignment-Based Versus Alignment-Free Sequence Comparison

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 [3]. Some computationally intensive multiple alignment methods align sequences strictly based on the order in which they receive them. 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 [11]. This implies that they can be more accurately aligned, but may result in incorrect phylogeny. Alignment can provide an optimized tree if a recursive approach is followed; however, this will increase the complexity of the problem. If the differences among the lengths of sequences are very high, the alignment performance significantly impacts tree generation.

The use of dynamic programming in alignment makes computation more complicated, and iterative steps limit their utility for large datasets. Therefore, consistent efforts have been made in developing and improving multiple sequence alignment methods for supporting variable length sequences with high accuracy and also for aligning a larger number of sequences simultaneously. Because of the problems associated with alignment-based phylogeny the importance of alignment-free methods is apparent [3]. Hence, the alignment quality affects the relationship created in a phylogenetic tree based on the consideration discussed above.

4.1 Alignment-Free Methods for Sequence Analysis

Alignment-free methods proposed in recent years can be classified into various categories as shown in Fig. 1. These include k-tuple based on the word frequencies, methods that represent the sequence without using the word frequencies, i.e. compression algorithms probabilistic methods and information theory-based method. 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 the frequency vector of subsequence. The probabilistic methods represent the sequences using the transition matrix of a Markov chain [12] of a pre-specified order, and comparison of two sequences is done by finding the distance between two transition matrices. Graphical representation comprising 2D or 3D or even 20D methods provides an easy way to view, sort and compare various sequences. Graphical representation further helps in recognizing major characteristics among similar biological sequences.

As discussed k-tuple method uses k-words to characterize the compositional features of a sequence numerically. A biological sequence is numerically converted into a vector or a matrix composed of the word frequency. The k-word frequency pro-

vides a fast arithmetic speed and can be applied to full sequences. The problem with k-tuple is a big value of k that poses a challenge in the computing time and space, and k-word methods underestimate or even ignore the importance of its location. The string-based distance measure uses substring matches with k mismatches.

5 Application of Phylogenetics in Cancer Studies

Cancer research is considered one of the most significant areas in the medical community. Mutations in genomic sequences are responsible for cancer development and increased aggressiveness in patients [13, 14]. The combination of all such genes mutations, or progression pathways, across a population can be summarized in a phylogeny describing the different evolutionary pathways [9]. Application of the phylogenetic tree can be explored for finding similarities among breast cancer subtypes based on gene data [14, 15]. Discovery of genes associated in cancer subtype help researchers to map different pathways to classify cancer subtypes according to their mutations. Methods of phylogenetic tree inference have proliferated in cancer genome studies such as breast cancer [13]. Phylogenetic can capture important mutational events among different cancer types; a network approach can also capture tumour similarities.

It has been observed from the literature that in cancer disease, the driver genes change the cancer progression, and it even affects the participation of other genes thus generating gene interaction network. Phylogenetic methods can solve the problem of class prediction by using a classification tree. Phylogenetic methods give us a deeper understanding of biological heterogeneity among cancer subtype.

6 Conclusion

The research focuses on the various methods of sequence analysis to generate phylogenetic trees. The limitations associated with sequence alignment methods lead to the development of alignment-free sequence analysis. However, most of the existing alignment-free methods are unable to build an accurate tree so more refinement is required in alignment-free methods. The phylogenetic study is not limited to species evolution, but disease evolution as well. Extending phylogenetic to disease diagnosis can give birth to new treatment options and understanding its progression.

Acknowledgements The research is funded by Department of Science and Technology, Delhi, under the sanction number SR/WOS-A/ET-1015/2015.

References

1. Lam, T.-Y., Hon, C.-C., Tang, J.W.: Use of phylogenetics in the molecular epidemiology and evolutionary studies of viral infections. Crit. Rev. Clin. Lab. Sci. **47**(1), 5–49 (2010)
2. Moret, B.M.E., Warnow, T.: Reconstructing optimal phylogenetic trees : a challenge in experimental algorithmics. In: Experimental Algorithmics, LNCS, pp. 163–180 (2002)
3. Vinga, S.: Editorial: alignment-free methods in computational biology. Brief. Bioinform. **15**(3), 341–342 (2014)
4. Geetika, Hanmandlu, M., Gaur, D.: Analyzing DNA strings using information theory concepts. In: ICTCS-16, ACM Conference, Udaipur, no. 9 (2016)
5. Munjal, G., Hanmandlu, M., Saini, A., Gaur, D.: Modified k-Tuple method for the construction of phylogenetic trees. Trends Bioinform. **8**(3), 75–85 (2015)
6. Schwartz, R., Schäffer, A.A.: The evolution of tumour phylogenetics: principles and practice. Nat. Rev. Genet. **18**(4), 213–229 (2017)
7. Chan, R.H., Chan, T.H., Yeung, H.M., Wang, R.W.: On maximum entropy principle for sequence comparison. IEEE/ACM Trans. Comput. Biol. Bioinform. **9**(1), 79–87 (2012)
8. Needleman, W.: A general method applicable to the search for similarity in the amino acid sequence of two proteins. J. Mol. Biol. **1970**(48), 443–453 (1969)
9. Smith, T.F., Waterman, M.S.: Comparison of biosequences. Adv. Appl. Math. **2**(4), 482–489 (1981)
10. Alon, N., Chor, B., Pardi, F., Rapoport, A.: Approximate maximum parsimony and ancestral maximum likelihood. IEEE/ACM Trans. Comput. Biol. Bioinform. **7**(1), 1–7 (2010)
11. Burr, T.: Phylogenetic trees in bioinformatics. Curr. Bioinform. **5**(1), 40–52 (2010)
12. Bai, F., Xu, J., Liu, L.: Weighted relative entropy for phylogenetic tree based on 2-step Markov Model. Math. Biosci. **246**(1), 8–13 (2013)
13. Somarelli, J., Ware, K., Kostadinov, R., Robinson, J., Amri, H., Abu-Asab, M., Fourie, N., Diogo, R., Swofford, D., Townsend, J.: PhyloOncology: understanding cancer through phylogenetic analysis. Biochimica et Biophysica Acta (BBA)—Rev. Cancer **1867**(2), 101–108 (2017)
14. Desper, R., Khan, J., Schäffer, A.: Tumor classification using phylogenetic methods on expression data. J. Theor. Biol. **228**(4), 477–496 (2004)
15. Munjal, G., Hanmandlu, M., Srivastava, S.: Novel gene selection method for breast cancer classification. J. Biochem. Technol. **8**(4), 1116–1120
16. Borozan, I., Watt, S., Ferretti, V.: Sequence analysis Integrating alignment-based and alignment-free sequence similarity measures for biological sequence classification. Bioinformatics **31**(January), 1396–1404 (2015)
17. Nemeth, C.: Hidden Markov models with applications to DNA sequence analysis. STOR-i, Lancaster University
18. Hohl, M., Ragan, M.A.: Is multiple-sequence alignment required for accurate inference of phylogeny? Syst. Biol. **56**(2), 206–221 (2007)
19. Potter, R.M.: Constructing phylogenetic trees using multiple sequence alignment. University of Washington (2008)
20. Burr, T.: Phylogenetic trees in bioinformatics. Curr. Bioinform. **5**(1), 40–52 (2010)
21. Cho, A.: Constructing phylogenetic trees using maximum likelihood. Ph.D. Thesis, Scripps women's college Claremont (2012)
22. Felsenstein, J.: PHYLIP. University of Washington Seattle, WA (1993)
23. Sardaraz, M., Tahir, M., Aziz Ikram, T., Bajwa, H.: Applications and algorithms for inference of huge phylogenetic trees: a review. Am. J. Bioinform. Res. **2**(1), 21–26 (2012)
24. Dawyndt, P., De Meyer, H., De Baets, B.: UPGMA clustering revisited: a weight-driven approach to transitive approximation. Int. J. Approx. Reason. **42**(3), 174–191 (2006)
25. Potiny, S.: An improved phylogenetic tree comparison method. Thesis University of North Carolina (2010)
26. Bryant, D., Moulton, V.: Neighbor-net: an agglomerative method for the construction of phylogenetic networks. Mol. Biol. Col **21**(2), 255–265 (2004)

27. Brinkman, F.S.L.: Bioinformatics: a practical guide to the analysis of genes and proteins. Publisher John Wiley and Sons (2001)
28. Munjal, G., Sharma, P., Gaur, D.: Sequence similarity using composition method. Int. J. Data Sci. **3**(1), 19–28. https://doi.org/10.1504/IJDS.2018.090626
29. Leimeister, C., Morgenstern, B.: Sequence analysis kmacs: the k-mismatch average common substring approach to alignment-free sequence comparison. Bionformatics **30**(14), 2000–2008 (2014)