Computational Methods for Phylogenetic Analysis

Student: Mohd Abdul Hai Zahid

Supervisors: Dr. R. C. Joshi and Dr. Ankush Mittal

Phylogenetics is the study of relationship among species or genes with the combination of molecular biology and mathematics. Most of the present phylogenetic analysis softwares and algorithms have limitations of low accuracy, restricting assumptions on size of the dataset, high time complexity, complex results which are difficult to interpret and several others which inhibits their widespread use by the researchers. In this work, we address several problems of phylogenetic analysis and propose better methods addressing prominent issues.

It is well known that the network representation of the evolutionary relationship provides a better understanding of the evolutionary process and the non-tree like events such as horizontal gene transfer, hybridization, recombination and homoplasy. A pattern recognition based sequence alignment algorithm is proposed which not only employs the similarity of SNP sites, as is generally done, but also the dissimilarity for the classification of the nodes into mutation and recombination nodes. Unlike the existing algorithms [1, 2, 3, 4, 5, 6], the proposed algorithm [7] conducts a row-based search to detect the recombination nodes. The existing algorithms search the columns for the detection of recombination. The number of columns

in a sequence may be far greater than the rows, which results in increased complexity of the previous algorithms.

Most of the individual researchers and research teams are concentrating on the evolutionary pathways of specific phylogenetic groups. Many efficient phylogenetic reconstruction methods, such as Maximum Parsimony [8] and Maximum Likelihood [9], are available. However, these methods lead to hard optimization problems and are limited to small number of taxa. Using these methods, more accurate phylogenetic trees can be constructed for small number of taxa in a reasonable time frame. On the other hand, distance-based [10, 11, 12] phylogenetic reconstruction methods can be used for large taxa, but the conversion from sequence data to distance data leads to loss of information. A method which exploits the features of distance and character based phylogenetic reconstruction methods is proposed in this thesis to construct phylogenetic supertree. The smaller trees can be constructed using accurate maximum parsimony or likelihood methods and then these smaller tree can be combined into single tree using distance based methods. We developed a variant of well known Unweighted-Pair-Group Method with Arithmetic mean (UPGMA) [11] for constructing the rooted supertree [13]. We also consider the problem of supertree reconstruction for unrooted trees input trees [14].

Most of the existing supertree methods combine the input trees based on the topological information carried by each of the input tree. Other evolutionary information such as fossil data, molecular dating data and actual divergence time estimates are usually ignored [15]. If the available evolutionary information is considered with tree topology for amalgamating the input collection of trees, the resulting supertree would be more accurate and resolved than the supertree constructed without using the additional infor-

mation. In this thesis, we propose a novel supertree method [16], which incorporates relative time divergence information with tree topologies. This method returns a tree even for incompatibilities such as divergence dates conflicts and incompatibility between topology and divergence dates.

Generally, most of the existing supertree methods are developed based on the implicit assumption that only leaf nodes are labelled in the input tree collection. In that case even the resulting supertree is also leaf labelled tree. On the other hand, the phylogenies constructed based on morphological studies often contain the labelled internal nodes, thus asking for a more generalized supertree approach. For example, TreeBase [17] maintains the database of published phylogenetic trees on different biological groups, and hence contains the tree representing the evolutionary relationship at different taxonomic levels [18]. In this work we propose an optimization based divide and conquer method [19] to combine semi-labelled trees. The proposed method returns a supertree even for (descendent level) incompatible input trees. Moreover, it also preserves all the nestings present in the input tree collection.

The exponential growth in the scientific literature makes it difficult for a researcher to navigate quickly through the desired information. Abbreviations and definitions are very important for understanding any scientific literature and new researchers often struggle to extract them from huge corpus. The experts in the domain may be interested in knowing the details of the individual documents. In this thesis, we propose a machine learning based phylogenetic question answering system to help answer the queries from the research communities. It focuses on answering natural language questions regarding phylogenetic algorithms.

The methods proposed above are supported by proofs and experimental

results are demonstrated over significant datasets to show the superiority and effectiveness of these methods. The aforementioned five contributions of this thesis provide an easy to use and generic platform for quick phylogenetic analysis in real-world applications.

References

[1] M. T. Hallet and J. lagergren. Effieceint algorithms for lateral gene transfer problems. pages 149–156, 2001. In proceedings of 5th international conference on computational molecular biology (RECOMB 01).

[2] V. Makarenkov and P. Legendre. From a phylogenetic tree to a reticulated network. *J. Comput Biol.*, 11:195–212, 2004.

[3] D. Bryant and V. Moulton. Neighbor-net: an agglomerative method for the construction of phylogenetic networks. *Mol. Biol. Evol.*, 21:255–265, 2004.

[4] J. Hein. Reconstructing evolution of sequences subject to recombination using parsimony. *Math. Biosci.*, 98:185–200, 1990.

[5] D. H. Huson. Splitstree: A program for analyzing and visualizing evolutionary data. *Bioinformatics*, 141:68–73, 1998.

[6] E. Satish D. Gusfield and C. Langley. Optimal efficient reconstruction of phylogenetic network with constrained recombination. *Journal of bioinformatics and computational biology*, 2:173–213, 2004.

[7] M. A. H. Zahid, A. Mittal, and R. C. Joshi. A pattern recognition based approach for phylogenetic network construction with constrained recombination. *Pattern Recognition*. In press.

[8] W. Fitch. Toward defining the course of evolution: minimum change for a specified tree topology. *Syst. Zool.*, 20:406–416, 1971.

[9] J. Felsenstein. Evolutionary trees from dna sequences: A maximum likelihood approach. *J. Mol. Evol.*, 17:368–376, 1981.

[10] W. M. Fitch and E. Margoliash. A non-sequential method for constructing trees and hierarchical classifications. *Journal of Molecular Evolution*, 18:30–37, 1967.

[11] R. R. Sokal and C.D. Michener. A statistical method for evaluating systematic relationships. *Univ. Kansas Sci. Bull.*, 28:1409–1438, 1958.

[12] Nei M. Saitou, N. The neighor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4:405–425, 1987.

[13] M. A. H. Zahid, A. Mittal, and R. C. Joshi. Least common ancestor based efficient method for constructing rooted supertrees. *Journal of Bioinformatics and Biomedical Engineering*, 1:1–6, 2005.

[14] M. A. H. Zahid, A. Mittal, and R. C. Joshi. A heuristic algorithm for optimal agreement supertrees. pages 595–599, 2005. In the proceedings of International Conference on Systemics, Cybernetics and Informatics (ICSCI 05), Hyderabad, India.

[15] D. Bryant, C. Semple, and M. Steel. supertree methods for ancestral time divergence date and other applications. In O. Bininda-emonds, editor, *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, pages 129–150. Kluwer, 2004.

[16] M. A. H. Zahid, A. Mittal, and R. C. Joshi. A supertree method for combining rooted phylogenetic trees with ancestral divergence time. pages 626:1–4, 2005. In the proceedings of 12th International Conference on BioMedical Engineering, IFMBE Proceedings, Singapore.

[17] www.treebase.org. Last accessed: 25 August 2006.

[18] R. D. M. Page. Taxonomy, supertrees, and the tree of life. In O. Bininda-emonds, editor, *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, pages 247–265. Kluwer, 2004.

[19] M. A. H. Zahid, A. Mittal, and R. C. Joshi. An optimization based approach for combining semi-labeled rooted phylogenetic trees. pages 624:1–4, 2005. In the proceedings of 12th International Conference on BioMedical Engineering, IFMBE Proceedings, Singapore.