International Journal of Engineering, Mathematical and Physical Sciences
ISSN: 2517-9934
Vol:2, No:6, 2008

Enhanced Character Based Algorithm for Small Parsimony

Parvinder Singh Sandhu, Sumeet Kaur Sehra, and Karmjit Kaur

***Abstract*—**Phylogenetic tree is a graphical representation of the evolutionary relationship among three or more genes or organisms. These trees show relatedness of data sets, species or genes divergence time and nature of their common ancestors. Quality of a phylogenetic tree requires parsimony criterion. Various approaches have been proposed for constructing most parsimonious trees. This paper is concerned about calculating and optimizing the changes of state that are needed called Small Parsimony Algorithms. This paper has proposed enhanced small parsimony algorithm to give better score based on number of evolutionary changes needed to produce the observed sequence changes tree and also give the ancestor of the given input.

***Keywords*—**Phylogenetic Analysis, Small Parsimony, Enhanced Fitch Algorithm, Enhanced Sakoff Algorithm.

I. Introduction

PHYLOGENETICS analysis plays major role in the interpretation of information on all characteristics of organisms, from structure and physiology to genomic. With the technological advances and the increasing availability of molecular data, their accurate reconstruction seems more attainable than ever. Phylogeny reflects the history of transmission of life's genetic information, and hence organizes knowledge of diverse organisms, genomes, and molecules. A reconstructed phylogeny provides invaluable information for resolving various issues. At the species level, a phylogeny provides characteristics of various species.

Phylogenetic Analysis is the main tool for representing evolutionary relationships among biological entities at the level of species. Since the evolutionary history is at best partially known, biologists, mathematicians, and computer scientists have designed a variety of criteria and methods for their accurate reconstruction. But there is always some space to increase the accuracy of reconstruction.

Basic Algorithms are less accurate and does not find the ancestor of the sequence. In simple sankoff algorithm, different values are given in transition and transversion. It does not consider different substitution for different values. In the present method, accuracy of small parsimony algorithms has been increased using Enhanced Fitch and Enhanced Sankoff algorithm. ESA considers different rates for different substitutions. It uses general 12-parameter model to calculate different rate of substitutions. Using different values at different nucleotide substitution would increase accuracy.

The remainder of this paper can be described as follows: Next section contains a description of the used algorithms for character based phylogenetic analysis. In Section III the proposed system, enhanced algorithms are described in detail. Sections IV provide description and results of experiments carried out. The paper ends with conclusions.

Parvinder S. Sandhu is Professor with Computer Science & Engineering Department, Rayat & Bahra Institute of Engineering & Bio-Technology, Sahauran, Distt. Mohali, Punjab-140104, India (phone: +91-98555-32004; e-mail: parvinder.sandhu@gmail.com).

S. K.Sehra is lecturer in Department of Computer Science & Engineering in Guru Nanak Dev Engg, College, Ludhiana ,Punjab, India (e-mail: sumeetksehra@gmail.com).

K. Kaur is with Electronics Engineering Department, Doaba Women Institute of Engineering and Technology, Kharar, Punjab, India (e-mail: karamsran@gmail.com).

II. Previous Work

There are many possibilities to reconstruct a phylogenetic tree from a set of objects. Purpose is to find the 'best' tree, or at least a good one. Judging the quality of a phylogenetic tree requires parsimony criterion. The general idea is to find the tree with the minimum amount of evolution, specifically, with the fewest number of evolutionary events. This tree is called the most parsimonious tree. There are several methods to reconstruct the most parsimonious tree from a set of data. First, is to find a possible tree. Second, is to calculate and optimize the changes of state that are needed. First problem is treated as large parsimony problem, and second one is small parsimony. Main purpose of small parsimony is to find the score of a given tree [8]. These algorithms are described in following paragraphs.

A. Fitch Algorithm

Walter Fitch published a dynamic programming algorithm that solves the small parsimony problem efficiently. Fitch's algorithm works on each set of states, as long as each state can change into each other. First the leaves of the tree are labeled with the current states. Then, the Fitch's set for each of the interior nodes of the tree is computed. To compute the Fitch's sets, the labeling of the two descendant nodes has to be considered. If they don't have an intersection, Fitch's set is the union and a penalty of 1 is added, otherwise it's simply the intersection and no penalty is added [7]. This is done until the root is reached. Fitch's set at the root is now the set of characters for which one can get a parsimonious labeling and the overall penalty is the number of changes needed in the tree.

Input: A phylogenetic tree T, with n leaves, and a single character c with a set of k possible values. Denote the value of the character for node v by c (v).Score = 0.

For each leaf v: $S(v) = \{c(v), l(v) = 0\}$ (1)

For each inner node v with children u, w

$$S(v)= \begin{Bmatrix} \{s(u) \cap s(w), s(u) \cap s(w) \neq \emptyset\} \\ \{s(u) \cup s(w), s(u) \cap s(w) \neq \emptyset\} \end{Bmatrix} \quad (2)$$

$$S(u) \cap S(w) \neq \emptyset . \quad (3)$$

$$S(u) \cap S(w) = \emptyset . \quad (4)$$

$$score=score+1. \quad (5)$$

To compute $S(v)$ and score the tree is traversed in post order- starting with the leaves and working our way down to the root. The parsimony score is then given by score. In total, the algorithm requires $O(nL)$ steps.

B. Sankoff Algorithm

Sankoff's algorithm is more complex than Fitch's, but it has the advantage that the changes from one state into another state can be weighted. In Sankoff algorithm cost matrix is used, in which the cost for changing from state i to state j is denoted by c_{ij} [1]. As in Fitch's algorithm, every node is considered in all trees; one tree for every state, but this time one doesn't have character sets but arrays with a cell for each possible state. The leafs will be labeled with arrays containing a 0 for an observed state and 1 otherwise. For a node a, the values of each cell in the array are computed with

$$S_a(i) = \min_j [c_{ij} + S_{l(j)}] + \min_k [c_{ik} + Sr_{(k)}] \quad (6)$$

In this equation, Sa is the actual node in state i, $S_{l(j)}$ is the left descendant in state j and $Sr_{(k)}$ is the right descendant in state k. This means that one has to search for the smallest possible cost for node a in state i by adding the minima for changing from state j to i plus the penalty at node l in state j and for changing from state k to i plus the penalty at node r in state k.[8]. Computing the arrays at the interior nodes of the tree a post order traversal is done until root is reached. At the root the minimum of the values is chosen in the array, this means $S = \min_i S_0(i)$, Now, S is the minimum number of evolutionary changes for this tree.

III. Proposed Work

A. Enhanced Fitch Algorithm

Enhanced Fitch algorithm uses Transition and transversion parameter to compute parsimony score, so with this accuracy increases. Fitch algorithm is mainly used for small data sets and provides fast speed of processing but its accuracy is not so fine, but adding this feature, algorithm accuracy will increase to some extent, which is the main feature of it. Secondly fitch algorithm does not provide common ancestor.

Common ancestor gives good correlation between different characters with little bias. Methods of ancestor reconstruction were important tools for evolutionary inference that are difficult to test empirically because ancestral states were rarely known with certainty. When the ancestor of a tree was not known then uncertain conditions or error occurred [4]. The fitch algorithm does not use the concept of transition-transversion.

In transition- transversion, nucleotides are divided in to two separate categories on the bases of the structure of their nitrogenous bases. [5] G and A are called purine because their nitrogenous bases have a two ring structure. In contrast, pyrimidines like C, T all have nitrogenous bases with only a one ring structure. When a purine with a purine or a pyrimidine with a pyrimidine is changed, then it is said to be transition and when a purine changes with a pyrimidine and vice versa, then it is said to be transversion.

TABLE I
Transition and Transversion

Transition	Transversion
A-G	A-C, C-A
G-A	G-T, T-G
C-T	A-T, T-A
T-C	G-C, C-G

The score allocation will be different for transition and transversion. By using this feature the accuracy of the algorithm for finding the score will increase. The algorithm is given as under, by using a newick format, it will find the score as well as ancestor of phylogenetic tree.

Input: A phylogenetic tree T, with n leaves, and a single character c with a set of k possible values. Denote the value of the character for node v by c (v).Score = 0.

For each leaf v: $S(v) = \{c(v), l(v) = 0\}$

For each inner node v with children u, w

$$S(v)= \begin{Bmatrix} \{s(u) \cap s(w), s(u) \cap s(w) \neq \emptyset\} \\ \{s(u) \cup s(w), s(u) \cap s(w) \neq \emptyset\} \end{Bmatrix}$$

$S(u) \cap S(w) = \emptyset$, $S(u) \cap S(w) = \emptyset$, $score=score+1$

$$S(v)= \begin{Bmatrix} Transition, Score = Score + 1 \\ Transversion, Score = Score + 2 \end{Bmatrix} \quad (7)$$

B. Enhanced Sankoff Algorithm

Simple sankoff algorithm uses 0 and 1based feature and is assumption based. It does not use different substitutions for each and every transition. ESA uses different rate of substitution for different transitions, by using 12 general parameter models. This model states that each every transaction must be treated differently. General 12-parameter model is being applied to get different substitution value for each and every transaction on the given tree.

In total there are 12 distinct neighbors independent substitution processes of a single nucleotides by another; four of them are so-called transitions that interchange a purine with

a purine or a pyrimidine with a pyrimidine. The remaining eight processes are the so-called transversions that interchange a purine with a pyrimidine and vice versa. The rates of these processes, $\alpha \rightarrow \beta$, will be denoted $r_{\alpha\beta}$, where $\alpha, \beta \in$ {A, C, G, T} denote a nucleotide. The model is parameterized by the substitution rates and the length of the time span, dt, the respective substitution processes acted upon the sequence, which would be the time between the observation of an ancestral sequence and its daughter sequence, T. time span is dt = 1 and with this choice the substitution rates are equal to the substitution frequencies giving the number of nucleotide substitutions per bp. Model includes neighbor independent processes only and is parameterized by 12 substitution frequencies. The set of all substitution frequencies will be denoted by {r}. Probability formulas are used to compute the different rates for different nucleotide substitution is

$$\sum \alpha\beta\gamma = \text{prob. } \alpha(\beta\gamma) + \text{prob. } \gamma(\alpha\beta) + \text{prob. } ß(\alpha\gamma) + \text{prob. } \alpha\, ß(\gamma) + \text{prob. } ß\gamma(\alpha) \quad (8)$$

According to the formulae first three transactions on the right hand side only considers substitutions of bracket values specifically of two values but in case of last two statements all the three values are being considered at the time of substitution [1]. Specifically first three values in right hand side describe first substitution and next one shows the substitution of others in phylogenetic tree. i.e. first three terms describe single nucleotide substitutions on the three sites [2] whereas the last two sums represent the neighbor dependent processes at the other sites.

$$(\alpha, \beta, \gamma : t = 0) = \begin{matrix} 1 if\ (\alpha, \beta, \gamma) = (\alpha 1, \alpha 2, \alpha 3) \\ 0 otherwise \end{matrix} \quad (9)$$

After applying these formulas, sankoff algorithm is applied to phylogenetic tree, which yielded score with higher accuracy [6].

IV. Results

The overall accuracy of a phylogenetic tree is often measured as the number of correct taxon bipartitions found on the estimated tree divided by the total number of taxon bipartitions possible for taxa [6]. Addition of taxa can break up long branches and help the parsimony method to become consistent. The overall accuracy calculated for different algorithms for newick format of tree, ((T, A), (C, (A, G))) of tree in Fig. 1.

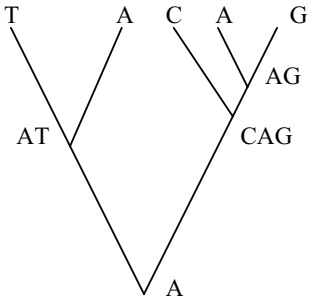

Fig. 1 Phylogenetic Tree

The accuracy is calculated using the formulae:

No. of taxon found on estimated tree / Total No. of taxons

A. Accuracy of Enhanced Fitch Algorithm

For EFA considers the two cases while calculating the score, so at the time of calculating the accuracy these features has to be included. These features are of transition and transversion. When there is transition, then value of 1 is used and when there is transversion then value of 2 is used

Number of taxon found on estimated tree is ={T, A, A, G}.
=[(T,A), (A,G)], [(T,A), (A,G)], [(T,G), (A,A)]
=2+1+2+1+2+0 =8

Total number of taxons= (No. of genes + No. of Transitions)
= (5+ {[T, A]+[C, A]+[A, G]+[C, G]})
= (5+ {2+2+2+1})
= 12

So, *accuracy* = No. of taxon found on estimated tree/ Total no. of taxons
=8/12
=0.67

Accuracy in percentage is: 0.67*100=67%

B. Accuracy of Enhanced Sankoff Algorithm

In case of enhanced sankoff algorithm, each transition must have different values and these values are being given to it, by using general 12 parameter model. Accuracy is being find out by using the same above stated formula for same phylogenetic tree.

S	A	T	G	C
A	0	1	0	1
T	3	0	3	4
G	1	1	0	2
C	2	2	2	0

Formula to calculate the accuracy is: No of taxon found on estimated tree/ Total no. of taxons.

So, here

Number of taxon found on estimated tree is ={T, A, A, G}.
=[(T,A), (A,G)], [(T,A), (A,G)], [(T,G), (A,A)]
=3+0+3+0+3+0
=9

Total number of taxons= (No. of genes + No. of Transitions)
= (5+ {[T, A]+[C, A]+[A, G]+[C, G]})
= (5+ {3+2+0+2})
= 12

So, accuracy= No. of taxon found on estimated tree/ Total no. of taxons =9/12
=0.75

Accuracy in percentage is: 0.75*100=75%.

V. Conclusion

The proposal this paper is to provide a enhanced algorithm in terms of finding the score of given tree i.e small parsimony algorithm. Both enhanced algorithms are giving better results are compared to their basic models. These algorithms would help the maximum parsimony to find the best tree in much accurate method and hence results in accurate judgment of evolutionary relationships among biological entities at the

level of species.

REFERENCES

[1] D. Huson "Algorithms in Bioinformatics", Syst. Zoology. 20:406-416., 2007.
[2] F. Arndt Peter and Hwa Terence "Identification and measurement of neighbor-dependent nucleotide substitution processes". Bioinformatics Original Paper Vol. 21 No. 10, Pages 2322–2328., 2005.
[3] F. Arndt Peter December 2006 "Reconstruction of ancestral nucleotide sequences and estimation of substitution frequencies in a star phylogeny". Gene 390 (2007) 75–83., 2006.
[4] H. Oakley Todd And W. Cunningham Clifford () "Independent Contrasts Succeed Where Ancestor Reconstruction Fails In A Known Bacteriophage Phylogeny". Evolution, 54(2), 2000, pp. 397–405., 2000.
[5] K. A. K. Strandberg and A. Salter Laura "A comparison of methods for estimating the transition:transversion ratio from DNA sequences". Department of Mathematics and Statistics, University of New Mexico, Albuquerque, NM 87131, USA. Molecular Phylogenetics and Evolution 32 (2004) 495–503. , 2004.
[6] R. Bruce, P. John Huelsenbeck, Yang Ziheng And Nielsen Rasmus "Taxon Sampling and the Accuracy of Large Phylogenies". Department of Ecology and Evolution, State University of New York, Stony Brook, NY 11794-5245, USA. Syst. Biol. 47 (4), 710–718, 1998.
[7] R. Fredrik "Parsimony: Counting Changes" (2) 73-86., 2005.
[8] S.Wing-Kin, Weiwei Cheng, Liang Yang "Combinatorial methods in bioinformatics: Phylogenetic Trees Reconstruction". Vol-2, Page-7-17,. 2004.

Parvinder S. Sandhu is currently working as Professor and Chair, Department of Computer Science and Engineering at Rayat-Bahra Institute of Engineering and Bio-Technology, Mohali, India. Earlier, he worked as Chair, Department of Computer Science & Engineering at Guru Nanak Dev Engineering College, Ludhiana (Punjab). He received his Master of Engineering degree in Software Engineering from Thapar University, Patiala (India) and Ph.D. from Guru Nanak Dev University in 2008. He is a member of ISTE and acted as member of BOS (Punjab Technical University, India). He has published 19 research papers in referred International Journals and 20 research papers in renowned International Conferences. His current research interests are Software Reusability, Software Cost Estimation, Software Maintenance and Bio-Informatics.

Sumeet Kaur Sehra is currently working as Lecturer in the Deptt. of Computer Science Engg. at Guru Nanak Dev Engg. College, Ludhiana, India. She has received her M.Tech in Comuter Science Engg. in 2006 from Punjab Agriculture University, Ludhiana. Her active areas of interests are Soft Computing, Bioinformatics and Artificial Intelligence.

Karamjit Kaur is currently working with Electronics Engineering Department of Doaba Women Institute of Engineering and Technology, Kharar, Punjab. She has received her bachelor's degree from Sant Longowal Institute Engg and Tech, Longowal, Sangrur in 2001. Currently pursuing master's in Electronics and Communication Engg. from Guru Nanak Dev Engg College, Ludhiana