International Journal of Biological, Life and Agricultural Sciences
ISSN: 2415-6612
Vol:1, No:10, 2007

A Maximum Parsimony Model to Reconstruct Phylogenetic Network in Honey Bee Evolution

Usha Chouhan and K. R. Pardasani

***Abstract*—**Phylogenies ; The evolutionary histories of groups of species are one of the most widely used tools throughout the life sciences, as well as objects of research with in systematic, evolutionary biology. In every phylogenetic analysis reconstruction produces trees. These trees represent the evolutionary histories of many groups of organisms, bacteria due to horizontal gene transfer and plants due to process of hybridization. The process of gene transfer in bacteria and hybridization in plants lead to reticulate networks, therefore, the methods of constructing trees fail in constructing reticulate networks. In this paper a model has been employed to reconstruct phylogenetic network in honey bee. This network represents reticulate evolution in honey bee. The maximum parsimony approach has been used to obtain this reticulate network.

***Keywords*—**Hybridization, HGT, Reticulate networks, Recombination, Species, Parsimony.

I. Introduction

PHYLOGENIES are the main tool for representing evolutionary relationships among biological entities. The biologists, mathematicians, and computer scientists are working to design a variety of methods for their reconstruction. Almost all such methods, however construct trees; yet scientists have long recognized that trees oversimplify our view of evolution science. They cannot take into account such events as hybrid speciation, horizontal gene transfer, recombination and gene conversion [7], [8]. These nontree events, usually called reticulations, give rise to edges that connect nodes from different branches of a tree, creating a directed acyclic graph structure that is usually called a phylogenetic network [1], [2]. To date, no accepted methodology for network reconstruction has been proposed. Various scientists have studied closely related problems, such as the compatibility of tree splits and to other indications that a tree structure is inadequate for the data at hand for detection and identification of horizontal gene transfer and, more generally detection and identification of recombination events in a number of biological studies of reticulation[11],14],[15].
In this paper, we describe an algorithm for modeling reticulate phylogenetic relationship among species to reconstruct phylogenetic network in honey bee by means of reticulated networks (RNs).The parsimony method has been studied and used extensively for phylogenetic trees. It is based on a minimum – information principle; in absence of information to the contrary, the best explanation for the observed data is that it involves the smallest number of manipulations or in the case of evolutionary histories it represents the fewest evolutionary events. As [5], [6] pointed out, parsimony can be extended to phylogenetic networks and it is observed that each individual site in a set of sequences labeling a network evolves down a tree contained in the network (i.e. a tree whose edges are edges of the network) In consequence the obvious extension is to define the parsimony score of a network as some overall sites of the parsimony score the best possible tree contained within the network for each site. But the parsimony method remains limited to just a few reticulations. If we generalize above view i.e. by adding reticulation events (in the form of additional edges), reconstructing maximum parsimony phylogenetic networks is NP-hard. As accessing the quality of the parsimony criteria for phylogenetic networks (rather than heuristics) and due to the absence of any efficient algorithms for solving the problem, we have to implement an exhaustive search method that traverses the entire space of network and considers the parsimony score of every network in the space. A version of the phylogenetic network reconstruction problem that applies to horizontal gene transfer and hybridization are explained as: Given an organismal (species) tree, compute an additional set of edges whose addition to the tree explains the horizontal gene transfer and hybridization events that occurred during the evolutionary history of the sequences [3]. Since these events are unknown and therefore the parsimony criterion is applied to seek the solution that is optimal with respect to this criterion.

Usha Chouhan is with the Maulana Azad National Institute of and Technology, Bhopal, MP-462051 INDIA (phone: 09425673321; fax: 91-755-2670562; e-mail: ycchouhan@gmail.com,ycchouhan@rediffmail.com).
K.R. Pardasani is with the Maulana Azad National Institute of and Technology, Bhopal, MP-462051 INDIA(phone: 09425358308; fax: 91-755-2670562; e-mail: kamalrajp@hotmail.com. kamalrajp@rediffmail.com).

II. METHOD & MATERIALS

When events such as horizontal gene transfer occurs the evolutionary history of set of organisms may not be modeled by phylogenetic trees; in this case, phylogenetic networks provide the correct model. In horizontal gene transfer (HGT), genetic material is transferred from one lineage to another; in an evolutionary scenario involving horizontal transfer. Certain sites (specified by a specific substring within the DNA sequence of the species into which the horizontal transferred DNA was inserted) are inherited through horizontal transfer from another species [12],[13].

A phylogenetic network N = (V, E) ,with a set X of n

leaves is a directed acyclic graph in which exactly one node, the root, has no incoming edges, and all other nodes have either one incoming edge tree nodes or two incoming edges reticulation nodes(see Fig.1a and b). In this paper, we focus on binary networks, i.e. networks in which the out degree of a tree node is two and out degree of a reticulated node is one. A tree T is contained (or induced) inside a network N if T can be obtained from N by removing exactly one of the two edges incoming into each reticulation node in N and using any applicable forced contractions denoted by T (N), the set of all trees induced by a network N. While a phylogenetic network models the evolutionary history of a set of organisms, the evolutionary histories of individual genes are trees which are contained inside the network [9],[16]. Reticulation events impose time constraints on the phylogenetic network. A phylogenetic network N = (V, E) defines a partial order on the set V of nodes. If we associate time t (u) with node u of N then, if there exists a directed path p from u to some other node v such that p contain at least one tree edge, we must have t (u) < t (v) in order to respect the time flow; moreover, if e = (u, v) is a network edge, then we must have t (u) = t (v), because hybridization is at the scale of evolution an instance as process given a network N, we say that p is a positive time directed path from u to v, if p is a directed path from u to v and p contains at least one tree edge. Given a network N, two nodes u and v cannot co-exist in time if there exists a sequence P = < P_1, P_2 ……. P_k> of paths such that: (I) P_i is a positive – time directed path, for every $l \leq I \leq k$ (ii) u is the tail of p_1 and v is the head of P_k, and (iii) for every $l \leq i \leq k$ – 1, there exists a network node whose two parents are the head of Pi and the tail of Pi+1 .Since events such as horizontal gene transfer occur between two lineages (nodes in the networks) that co- exist in time, a phylogenetic network N must satisfy the property : If two nodes x and y cannot co-exist in time then they cannot participate in a reticulation event, that is the network cannot include either of the two edges (x, y) and (y, x).

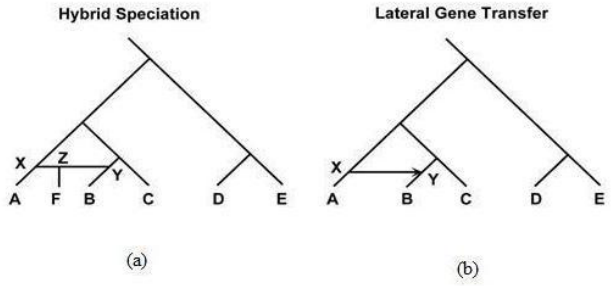

Fig. 1 (a) and (b) A phylogenetic network that consists of an organismal tree and an additional edge that corresponds to hybidization and HGT event.

Parsimony is one of the most popular methods used for phylogenetic tree reconstruction. Roughly this method is based on the assumption that "evolution is parsimonious'', i.e., the best evolutionary trees are the ones that minimize the number of changes along the edges of the tree. The evolution is parsimony formulated as given below:

The hamming distance between two equal length sequences x and y denoted by H(x,y) is the number of positions j such that $x_j \leq y_j$ .Given a fully labeled tree T, i.e., a tree in which each node v is labeled by a sequence Sv over some alphabet ∑, and define the hamming distance of an edge e€ E(T), denoted by H (e), to be H (S_u, V_s.) where u and v are the two endpoints of e and define the parsimony score of a tree T[4],[10].

Input : Set S of n aligned sequence of length k.

Output : A phylogenetic tree T is leaf – labeled by sequences in S and additional sequences of length k labeling the internal nodes of T such that ∑ H (i, j) is minimized.

The maximum parsimony score for a dataset of 4 nucleic-acid sequences is calculated. Consider the following set of homologous sequences:

	Site								
Sequence	1	2	3	4	5	6	7	8	9
1	A	A	G	A	G	T	G	C	A
2	A	G	C	C	G	T	G	C	G
3	A	G	A	T	A	T	C	C	A
4	A	G	A	G	A	T	C	C	G

For four OTUs there are three possible unrooted trees. The trees are then analysed by searching for the ancestral sequences and by counting the number of mutations required to explain the respective trees as shown below:

```
(1) AAGAGTGCA        AGATATCCA (3)
       \4       2/          Number of mutations
        \   4   /
     AGCCGTGCG --- AGAGATCCG        Tree (a):  11
        /       \
       /0       0\
(2) AGCCGTGCG        AGAGATCCG (4)
```

```
(1) AAGAGTGCA        AGCCGTGCG (2)
       \1       3/
        \   5   /
     AGGAGTGCA --- AGAGGTCCG        Tree (b):  14
        /       \
       /4       1\
(3) AGATATCCA        AGAGATCCG (4)
```

(1) AAGAGTGCA AGCCGTGCG (2)
\1 3/
\ 5 /
AGGAGTGCA --- AGATGTCCG Tree (c): 16
/ \
/5 2\
(4) AGAGATCCG AGATATCCA (3)

Tree (a) has the topology with the least number of mutations and thus is the most parsimonious tree. The same procedure apply four data set of honey bee and the result obtained in next section.

It works by finding the tree which can explain the observed sequences with a minimal number of substitutions. Instead of building a tree, it assigns a cost to a given tree, and it is necessary to search through all topologies, or to pursue a more efficient search strategy that achieves this effect in order to identify the best tree. The two main components of the algorithms are:
(i)the computation of a cost for a given tree T;
(ii) a search through all trees, to find the overall minimum of this cost.

The four nucleotide sequences and their aligned results of (honey bee) Apis florea, Apis dorsata,Apis cerana, Apis mellifera have been collected from NCBI Site.The ascension numbers of these sequences are given in Appendix.

From above sequences try out different trees for these four sequences and count number of substitutions needed in each tree to the ancestral nodes so to minimize the number of changes needed in the whole tree. In Fig. 2 we get the optional MP tree.

The parsimony score of a fully labeled tree T, is $\sum_{e \varepsilon E(T)} H(e)$.Given a set S of sequences, a maximum parsimony tree for set S is a tree leaf – labeled by S and assigned labels for the internal nodes, of minimum parsimony score. Given a set s of sequences, the parsimony problem is to find a maximum parsimony phylogenetic tree T for the set S The problem of computing the parsimony score of a fixed leaf – labeled tree is solvable in polynomial time. Parsimony on phylogenetic networks, the evolutionary history of a site i in a set S of sequence that evolved on a network, N is captured by one of the trees contained inside the network N.

Therefore, a natural way to extend true tree – based parsimony score to fit a data set that evolved on a network is to define the parsimony score of that site over all trees contained inside the network. This extension was first introduced by Hein at.el [5],[6] in the context of meiotic recombination and formalize general definition of parsimony. The parsimony score of a network N leaf – labeled by a set S of taxa, is N cost (N, S) = $\sum_{bi\varepsilon\beta}$ (min $_{T \varepsilon T(N)}$ T cost (T, b_i),where β is a set of blocks of equal length that partition the sequences, T cost (T, b_i) is the number of changes of block b_i on tree T, and T (N) denotes the set of trees contained inside network N..In above criterion, we would want to reconstruct a phylogenetic network whose parsimony score is minimized. In the case of horizontal gene transfer [1],[10], it is observed that the underlying organism tree is reconstructible Hence, the problem of reconstructing phylogenetic networks in this case becomes one of computing a set of edges whose addition to the organismal tree explains the horizontal gene.

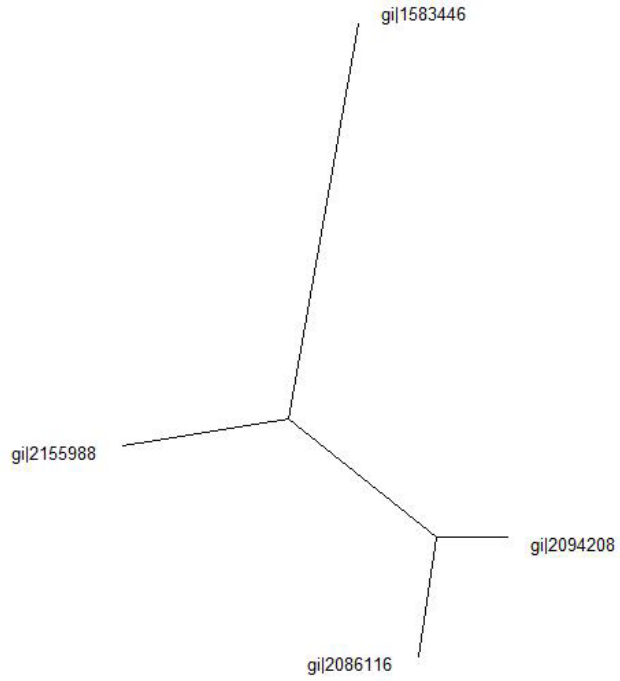

Fig. 2 Optional MP tree by tree view

III. RESULTS AND DISCUSION

Bees are a diverse, fascinating, and important group of insects with an intimate ecological interrelationship with the angiosperm (flowering) plants. The enormous radiation of the flowering plants may be due in part to the nearly simultaneous diversification of the bees. Today bees are one of the most economically and ecologically important insect groups. There are over 16,000 described species of bees and we are just beginning to understand the basal phylogeny of the bees, their historical biogeography, and the antiquity of bees. Bees are insects of the Order Hymenoptera which feed on pollen and nectar. They constitute a group of about 20 000 species throughout the world, known taxonomically as the Super family Apoidea. Honeybees of the genus Apis belong to the family Apidae, a sub-group of this super family. Although the question of how many honeybee species exist is still debated among taxonomists, at least four species are commonly recognized: the dwarf, or midget, bee Apis florea, the giant, or rock, bee Apis dorsata, the oriental (Indian, Chinese, Japanese, etc.) bee Apis cerana, and the common (European, African, etc.) honeybee Apis mellifera [7].

A study is performed for honey bee of four species Apis florea, Apis dorsata, Apis cerana, Apis mellifera, evolution using Maximum parsimony to generate phylogenetic tree on four taxa (shown in fig.3). This tree was produced by BioEdit, It is the tree that requires the least amount of mutation

(according to some measure) in order to explain the sequences that represent the leaves.

DNA parsimony algorithm, version 3.6a2.1
One most parsimonious tree found:

```
 +-----gi|2094208
 |
 |             +--------gi|2155988
 1-----------------2
 |             +--------gi|1583446
 |
 +----------gi|2086116
```

Fig. 3 parsimonious tree by BioEdit

Requires a total of 2002.000
Between and length

1	gi|2094208	0.116240
1	2	0.300586
2	gi|2155988	0.267106
2	gi|1583446	0.650898
1	gi|2086116	0.195751

The phylogeny clearly separated two groups of bees with the species *A. mellifera*, *A. dorsata* forming the first group and species *A. cerana, A. florea* the second group.

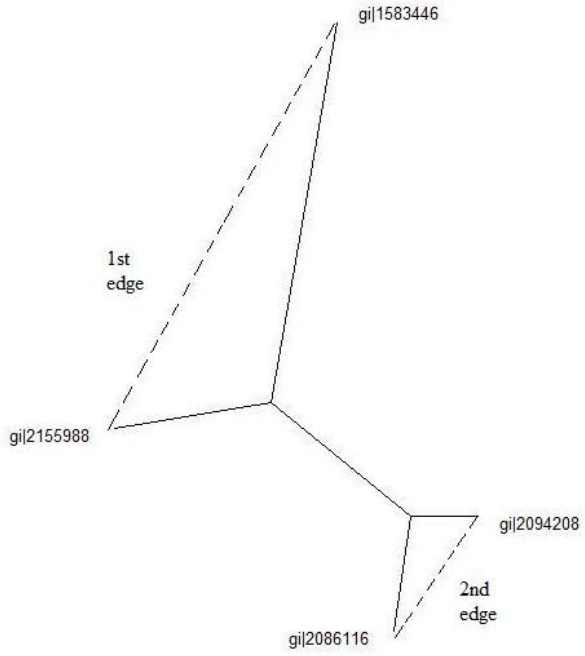

Fig. 4 Phylogenetic network

Parsimony infers phylogenetic networks adding one by one a reticulation edges and evaluating the possible mutations between sequences. The phylogenetic networks show that the two species are genetically closer to each other than it is represented by the phylogenetic tree. Fig. 4 depicts what may have happened during evolution: a recent ancestor of *A. cerana* may have hybridized or HGT with one of the recent ancestors of *A. florea* to produce the modern *A. florea* bee. Or, conversely, a recent ancestor of *A. florea* may have hybridized or HGT with one of the recent ancestors of *A. cerana* to produce the modern *A. cerana* species. This hypothesis is in agreement with the belief, based on biological and behavioral data, that *A. florea* and *A. cerana* have shared a close common ancestor in relatively recent times .The aim of parsimony methods is to find the phylogenetic network with minimum total length between sequences. That is the tree with the smallest number of evolutionary changes explaining the observed data in the network .These results indicate the relevance of the reticulogram model for the honeybee data, where reticulation branches bring to light conflicting features that are embedded in the phylogenetic tree. In Fig.4 the optimal parsimony scores are almost identical, regardless of the number of edges added, which implies that HGT events are inferred. Indeed, these 4 datasets evolved in their entirety down the organismal trees, and hence HGT events were present. Fig.4 shows the results on datasets whose evolution involves a single HGT event, between two closely related species Fig.4 show a much sharper decrease in the optimal parsimony score when adding the first edge, compared to the decrease in score when adding a second edge and so on. The contrast between the effects of adding the first and second edges is not as clear when the HGT event is between two closely related species, yet the decrease in the parsimony score after adding the first edge is not very large, which is a reflection of the hardness of detecting HGT events between closer organisms. In this case, the parsimony criterion may underestimate the number of HGT events. However, we predict that if both HGT events were between divergent organisms, we would see a sharper decrease in the parsimony score when adding the second edge.

APPENDIX

The ascension numbers of Nucleotide Sequences of Apis honey bee.

>gi|208611638|gb|FJ348345.1| Apis florea isolate India1 large subunit ribosomal RNA gene, partial sequence; mitochondrial.

>gi|158344660|gb|EU100935.1| Apis dorsata isolate DorsIII22Maly complementary sex determiner (csd)mRNA, complete cds

>gi|209420845|gb|FJ229480.1| Apis cerana cytochrome (cytb) gene, partial cds; mitochondrial.

>gi|215598870|ref|NM_001142461.1| Apis mellifera transmembrane protein 98 (Tmem98), mRNA.

ACKNOWLEDGMENT

The authors are highly grateful to Department of Biotechnology, New Delhi for providing support for this work under Bioinformatics Infrastructure Facility of DBT at MANIT Bhopal.

REFERENCES

[1] C. Dutta And A. Pan,"Horizontal gene transfer and bacterial diversity" ,*J. Biosci*. (*Suppl. 1*) 27 27–33, 2002.
[2] C.T Nguyen,B. Nguyen, W.K.Sung and L. Zhang,'Reconstructing Recombination Network from Sequence Data :The Small Parsimoney Problem",IEEE CS,Cl and EMB Societies & the ACM,Vol.4,No.3,pg 394-401,2007.
[3] D.H. Huson and D. Bryant,"Application of Phylogenetic Networks in Evolutionary Studies", Mol. Biol. Evol. 23(2):254–267. 2006.
[4] G. Jin, L. Nakhleh, S. Snir and T. Tulle,"Efficient parsimony based methods for phylogenetic network Reconstruction",Vol. 23 ECCB, pages e123–e128. 2006.
[5] J. Hein,"Reconstructing evolution of sequences subject to recombination using parsimon",Maths.Biosci.98, 185-200, 1990.
[6] J. Hein,"A heuristic method to reconstruct the history of sequences subject to recombination",J.mol.Evol.36,396-405, 1993.
[7] J. Archer, J. W. Pinney, J. Fan, E. S Loriere," Identifying the Important HIV-1 Recombination Breakpoints", PLoS Comput Biol. 4(9): e1000178, 2008.
[8] M.T Hallet and J. Lagergren,"Efficient algorithms for lateral gene transfer problems. In: Proceedings of the 5th Ann Int Conf Compt Mol Biol (RECOMB 01), New York, and ASM Press. pp 149-156, 2001.
[9] L. Nakhleh, D. Ruth's,"Gene Trees, species Trees, and species Networks", Lecture notes. RECOMB, San Diego, California USA, 1:1-24, 2003.
[10] L. Nakhleh,G.J.,F Zhao,"Reconstructing Phylogenetic Networks Using Maximum Parsimony", RECOMB, USA, 1:200-210, 2004.
[11] L. O. Martins,E.Leal,H. Kishino,"Phylogenetic Detection of Recombination with a Bayesian Prior on the Distance between Trees",PLoS One,Volume 3,Issue 7,July 2008,e2651.
[12] V. Makarenkov, P . Legendre and Y. Desdevises,"Modeling phylogenetic relationships using reticulated networks",Zool Scrip 33:89–96, 2004.
[13] V. Makarenkov,"T-Rex: reconstructing and visualizing phylogenetic trees and reticulation networks,",Bioinf 17:664–668, 2001.
[14] V. Makarenkov and P. Legendre ,"From a phylogenetic tree to a reticulated network", J Comput Biol 11:195–212, 2004.
[15] D. Posada and K.A Crandall,"Evaluation of methods for detecting recombination from DNA sequences: Computer simulations", Proc Natl Acad Sci USA 98(24):13757-13762, 2001.
[16] W.K Sung, "Phylogenetic Trees Reconstruction", CS5238: 1-7, 2005.