ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 3, Issue 1, July 2013

Particle Swarm Optimization for Hylogenetic Tree Reconstruction

Dr. P. Thirunavukarasu, A. Thanithamil
[1,2]Assistant Professor -P.G & Research Department of Mathematics, Periyar E.V.R College
Tiruchirappalli – 620 023, Tamilnadu, South India.

***Abstract*— In this paper we compare and introduce methodologies which can perform a highly accurate Phylogenetic analysis. The Multi-Stack algorithm categorically is a distance-based method. Thus it uses only the distance values of the sequences of interest to build a Phylogenetic tree. And also we propose a particle swarm optimization for maximum parsimony based Phylogenetic trees reconstruction.**

***Index Terms*— Phylogenetic tree, UPGMA, Neighbor Joining, Maximum parsimony, Particle swarm optimization.**

I. INTRODUCTION AND PRELIMINARIES

A Phylogenetic tree or evolutionary tree is a branching diagram or "tree" showing the inferred evolutionary relationships among various biological species or other entities based upon similarities and differences in their physical and/or genetic characteristics. The taxa joined together in the tree are implied to have descended from a common ancestor. In a rooted Phylogenetic tree, each node with descendants represents the inferred most recent common ancestor of the descendants and the edge lengths in some trees may be interpreted as time estimates. Each node is called a taxonomic unit. Internal nodes are generally called hypothetical taxonomic units (HTUs) as they cannot be directly observed. Trees are useful in fields of biology such as bioinformatics, systematic and comparative Phylogenetic. We construct Phylogenetic trees to illustrate the evolutionary relationships among a group of organisms. The purposes of Phylogenetic studies are reconstructing evolutionary ties between organisms and estimate the time of divergence between organisms since they last shared a common ancestor. There are several types of data that can be used to build Phylogenetic trees: Traditionally, Phylogenetic trees were built from morphological features (e.g., beak shapes, presence of feathers, number of legs, etc). Today, we use mostly molecular data like DNA sequences and protein sequences. The definition of Phylogenetic Trees defined in section II. Distance-based Approach is discussed in secion III. UPGMA and Neighbor Joining derived in section IV. Proposed PSO discussed in section V. Finally, the paper is concluded.

II. DEFINITION OF A PHYLOGENETIC TREE

The evolutionary history of biological objects (e.g. species, genes, proteins, genomes) is general can be represented as a tree structure, where each leaf corresponds to a biological object of interest. These biological objects will be denoted by X. The inner points of the tree can be regarded as hypothetical ancestors. In the terminology we follow, a Phylogenetic tree is a leaf-labeled tree.

Definition 2.1: A Phylogenetic tree is an ordered pair (T, ϕ), where T is a tree (acyclic connected graph) whose vertex set V(T) contains at most one element which is of degree two and $\phi : X \rightarrow L(T)$ is a bijection between the leaf set L(T) of T and the set X. The function ϕ is called the labeling function.

A Phylogenetic tree is rooted Phylogenetic tree if it has a vertex $r \in V$ which is of degree two. The vertex r is called as the root of the Phylogenetic tree.

Definition 2.2: If every non-root interior point of a rooted Phylogenetic tree T is of degree three, then T is called a binary Phylogenetic tree.

Definition 2.3: A Phylogenetic tree is a weighted Phylogenetic tree if there is a nonnegative real function on its edge set w : E(T) $\rightarrow R^+$ .The interior vertices of a Phylogenetic tree can be interpreted in many ways, depending on the type of the biological object studied. For example, if we reconstruct a tree for a set of genes, then each interior points can be interpreted as a so-called evolutionary event like a gene duplication or gene specialization. We should mention here that if RB(n) stands for the set of all rooted Phylogenetic tree over the set |x| = n, then |RB(n)| = (2n − 3)!!. When restricted to binary, rooted topologies (the kind of trees that biologists prefer to work with), the number of trees for a given set of taxa (n) is : (2n − 3)!!=

$$\frac{(2n-3)!!}{2^{n-1}(n-1)!} = 3X5X7X...(2n-3)$$

This means that the size of tree space grows at a super expononential rate with the seze of the texon set. This is why elementary approaches, such as eshaustive search, are hardly suitable even for a small taxon set.

Phylogenetic tree construction methods are widely accepted to fall into one of two categories: distance based and character based. These two categories both offer a vast variety of options when constructing trees in two different directions. The most common distance based methods are the unwieghted pair group method using arithmetic averages (UPGMA) [3], Neighbor Joining [4] and the Fitch and Margoliash [5] algorithms that are all based on the initial creation of a distance matrix. The alternative to these methods is the character based methods such as maximum

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 3, Issue 1, July 2013

parsimony [6] and maximum likelihood [7] which take a probabilistic approach to tree construction.

III. DISTANCE-BASED APPROACH

In distance-based approach, we shall assume that just the similarity distances between the biological objects are available. Thus, before we perform a phylogenetic analysis we need a distance function. Using this function we can express numerically the 'similarities' or even the 'dissimilarities' between the set of objects X. If we have a formal definition for a 'similarity' measure, which can express the similarities of sequences of fixed length, then a monotone decreasing transformation of the similarities can be used as a distance-like measure, and vice-versa. Several distance functions are now commonly used which define distance values between character strings over a given alphabet. In the course of evolution the changes in the sequences are called mutations. For the assessment of how two sequences are related from an evolutionary point of view, we need to identify these changes in the sequences. To do this the most commonly used method is the general alignment model, which in bioinformatics is also known as the Needleman-Wunsch algorithm. The method itself carries out a pair wise alignment for the sequences. Based on this alignment, we can determine the number of positions where two sequences are identical or differ from each other. After the alignment step, the distances of sequences are calculated based on a time continuous Markov Chain. These models determine the evolutionary distances of sequences based on the number and type of mutations which were identified during the pair wise alignment of sequences. This approach can take into account the case when in a certain position duplicated mutation takes place. For example, A $\rightarrow$ C occurs in one position, and then there is a C $\rightarrow$ A mutation again. Distance-based tree building method is the Neighbor-Joining method. The NJ method can be considered as a divisive method, namely it constructs the Phylogenetic tree via a bottom-up clustering procedure. The reasons for the success of the NJ method are being studied nowadays. The tree reconstruction methods outlined above belong to the class of agglomerative hierarchical clustering algorithms, since during each iteration they divide a cluster (top down) or join two clusters (bottom-up approach) based on a function defined on the clusters. Furthermore, the tree building methods assign edge weights to the edges of the output tree. Since the vertices of a Phylogenetic tree represent biological objects (the inner points represent only hypothetical objects), the edge lengths or edge weights represent evolutionary distances of objects. If we would like to compare the outputs of the distance-based methods, then which tree should we select? One way of answering this is by investigating the difference between the predefined distance values on X and the patristic distance defined by a Phylogenetic tree over X. Hence let us define the tree error e_T of a Phylogenetic tree T for a distance function d by the following formula:

$$e_T = \sum_{x,y \in X} \left(d(x,y) - d_T(x,y)\right)^2$$

Then the so-called path-edge incidence matrix P_T for a Phylogenetic tree T having n leaves is given by : PT(p, e) = 1, if e $e \in p$ $P_T(p,e) = 0$,if otherwise whose columns correspond to the edges of T, while the rows correspond to the paths between the leaves of T. Clearly this matrix has n − 1 columns and n rows, by solving an optimization problem we can obtain the minimal tree error, where the x vector contains the optimal edge weights of T, and the vector d contains the distance values according to the P_T topology matrix: $e_T = \min_{x \in R_+^{n-}} \left\| \left(P_T X - d\right) \right\|$ Based on the minimal tree errors the weighted Phylogenetic trees over a set X can be Ranked. Day showed that the problem of finding tree over a set X which has a minimal tree error is in general NP-complete [4]. This approach is used in the field of speech Recognition, and adopted this method to reconstruct Phylogenetic trees because the MS approach is very suitable for finding the heuristic solution of problems which have an enormous solution space.

IV. UPGMA AND NEIGHBOR JOINING

UPGMA and Neighbor Joining use a clustering procedure that is commonly found in data mining techniques. The method is simple and intuitive [8] which makes it appealing. The method works by clustering nodes at each stage and then forming a new node on a tree. This process continues from the bottom of the tree and in each step a new node is added, and the tree grows upward. The length of the branch at each step is determined by the difference in heights of the nodes at each end of the branch. UPGMA has built in assumptions that the tree is additive and that all nodes are equally distance from the root. Since a molecular clock hypothesis assumption poses biological issues, UPGMA is not used much today, but gave way to a very common approach now termed Neighbor Joining [4]. Assign each taxon to its own cluster define one leaf for each taxon; place it at height 0 while more than two clusters determine two clusters i, j with smallest d_{ij} define a new cluster $C_k = C_i \cup C_j$ define a node k with children i and j, place it at height $\frac{d_{ij}}{2}$ replace clusters i and j with k compute distance between k and other clusters join last two clusters, i and j by root at height $\frac{d_{ij}}{2}$ given a new cluster C_k formed by merging C_i and C_j. we can calculate the distance between C_k and any other cluster C_j as follows

$$d_{kl} = \frac{d_{il}|C_i| + d_{jl}|C_j|}{|C_i| + |C_j|}.$$

Neighbor Joining (NJ) works like UPGMA in that it creates a new distance matrix at each step, and creates the tree based on the matrices. The difference is that NJ does not construct clusters but directly calculates distances to internal nodes. The first step in the NJ algorithm is to create a matrix with the Hamming distance between each node or taxa. The minimal distance is then used to calculate the distance from

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 3, Issue 1, July 2013

the two nodes to the node that directly links them. From there, a new matrix is calculated and the new node is substituted for the original nodes that are now joined. The advantage here is that there is not an assumption about the distances between nodes since it is directly calculated.

V. NEIGHBOR JOINING ALGORITHM

Define the tree T = set of leaf nodes, L = T while more than two sub trees in T pick the pair i, j in L with minimal Dij add to T a new node k joining i and j determine new distances

$$d_{ik} = \frac{1}{2}\left(d_{ij} + r_i - r_j\right) \quad (4.1)$$

$$d_{jk} = d_{ij} - d - ik \quad (4.2)$$

$$d_{km} = \frac{1}{2}\left(d_{im} + d_{jm} - d_{ij}\right) \text{for all other m} \in \text{L;} \quad (4.3)$$

Remove i and j from L and insert k (treat it like a leaf) join two remaining subtrees, i and j with edge of length dij . Additivity is a measurement that depends on the distance measure used. Neighbor Joining works even if the lengths are not additive but the tree is no longer guaranteed to be the correct tree. There exists a four-point condition that can be used to test for additivity. A result of additivity is that the sum of the lengths of two distances must be greater than the third distance. For example, let dij represent the distance from i to j. If four nodes exist, then dij + dkl = dik + djl and is greater than dil + djk. This is because the inclusion of the link between the two smaller clusters is common in two of the distance sums. Since the Neighbor Joining approach to tree construction takes advantage of common clustering techniques. It produces an unrooted tree that shows the relationship between sequences without assigning a root node from which all other sequences have been derived. In order to construct a tree with a common ancestor node, an out-group species is chosen that is distantly related to the remaining sequences. The location where the new species connects to the recently constructed tree is a good indicator for the most likely location for the root of the tree. If it is not easy or possible to find an out-group, other strategies allow one to locate a root of the tree such as using the midpoint of the longest chain of consecutive edges, which would indicate the root if the tree followed the molecular clock within reason. The following Particle swarm optimization is a new approach to maximum parsimony based Phylogenetic trees reconstruction.

VI. PARTICLE SWARM OPTIMIZATION - [11], [12]

Swarm Intelligence (SI) is an innovative distributed intelligent paradigm for solving optimization problems that originally took its inspiration from the biological examples by swarming, flocking and herding phenomena in vertebrates. Particle Swarm Optimization (PSO) incorporates swarming behaviors observed in flocks of birds, schools of fish, or swarms of bees, and even human social behavior, from which the idea is emerged (Kennedy, 2001) (Clerc , 2002), (Parsopoulos, 2004). PSO is a population-based optimization tool, which could be implemented and applied easily to solve various function optimization problems. As an algorithm, the main strength of PSO is its fast convergence, The original PSO formulae define each particle as potential solution to a problem in D-dimensional space. The position of particle " i " is represented as

$$X_i=(x_{i1},x_{i2},.....x_{iD})$$

Each particle also maintains a memory of its previous best position, represented as

$$P_i=(p_{i1},p_{i2},.....p_{iD})$$

A particle in a swarm is moving; hence, it has a velocity, which can be represented as

$$V_i=(v_{i1},v_{i2},.....v_{iD})$$

Each particle knows its best value so far(pbest) and its position. Moreover, each particle knows the best value so far in the group (gbest) among pbests. This information is analogy of knowledge of how the other particles around them have performed.

Each particle tries to modify its position using the following information:

1. The distance between the current position and pbest
2. The distance between the current position and gbest.

This modification can be represented by the concept of velocity. Velocity of each agent can be modified by the following equation (5.1) in inertia weight approach (IWA)

$$v_{id} = w^* v_{id}+c_1^*r_1^*(P_{id} -X_{id})+c_2^*r_2^*(P_{gd} - X_{id}) \quad (5.1)$$

Where, v_{id} – velocity of particle
x_{id} – current position of particle
w – inertial factor
c_1- determine the relative influence of the cognitive component
c_2- determine the relative influence of the social component
P_{id}- pbest of particle i
P_{gd} – gbest of the group
r_1,r_2 – random numbers

Where w is called as the inertia factor which controls the influence of previous velocity on the new velocity, r_1 and r_2 are the random numbers, which are used to maintain the diversity of the population, and are uniformly distributed in the interval [0,1]. C_1 is a postive constance, called as coefficient of the self-recognition component, C_2 is a positive constant, called as coefficient of the social component. From equation (5.1), a particle decides where to move next, considering its own experience, which is the memory of its best past position, and the experience of its most successful particle in the swarm. In the particle swarm model, the particle searches the solutions in the problem space with a range [-s,s].

Simple PSO has the following steps:

1. Create initial generation P(0). Let t = 0.
2. For each individual i ∈ P(t), evaluate by (5.1).
3. Create generation P(t + 1) by PSO
4. Let t = t+1. Unless t equals the maximum number of Generations, return to Step 2.

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 3, Issue 1, July 2013

VII. CONCLUSION

The proposed particle swarm optimization (PSO) is a good candidate tool for new approach to maximum parsimony based tree reconstruction will support this endeavors, and provide new insights. This comparative based evaluation handles the calculation of the individuals. As the problem is NP-complete in nature, the use of PSO scheme proposed has produced good result for real data. The results of this study may lead to the development of effective PSO for solving other model of tree reconstruction.

REFERENCES

[1] Adams E. N., Consensus techniques and the comparison of taxonomic trees. Systematic Zoology, 21 (1972), 390-397.

[2] Altschul S. F., Gish W., Miller W., Myers E. W., and Lipman D. J., Basic local alignment search tool. J Mol Biol, 215(3) (October 1990), 403-410.

[3] Busa-Fekete R., Kocsor A., and Bagyinka Cs. A multi-stack based Phylogenetic tree building method. Lecture Notes in Bioinformatics, 4463 (2007), 49-60.

[4] Bryant D., A classication of consensus methods for phylogenetics. Bioconsensus, Discrete Mathematics and Theoretical Computer Science, 61 (2001), 163-184.

[5] Day W. H. E., Computational complexity of inferring phylogenies from dissimilarity matrices. Bulletin of Mathematical Biology, 49 (1986), 461-467.

[6] Eisen J. A., Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. Genome Res., 8 (1998), 163 - 7.

[7] Felsenstein J., Evolutionary trees from dna sequences: a maximum Likelihood approach. J Mol Evol, 17(6) (1981), 368-376.

[8] Gotoh O., An improved algorithm for matching biological sequences. J Mol Biol, 162(3) (December 1982), 705-708.

[9] Jukes T. H. and Cantor C. R., Evolution of protein molecules. Mammalian Protein Metabolism, Academic Press, New York, edited by H. N. MUNRO,(1969), 21-132.

[10] Kimura M., A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. Journal of Molecular Evolution. 16 (1980), 111-120.

[11] J. Kennedyand R. Eberhart, "Particle Swarm Optimization," IEEE Proceedings on 1st International Conference on Neural Networks (Perth, Australia), 1942-1948, 1995.

[12] J. Kennedy and R. Eberhart, "A Discrete Binary Version of the Particle Swarm Algorithm," IEEE Conference on Systems, Man, and Cybernetics, Orlando, FA, 4104-4109, 1997.

AUTHOR'S PROFILE

Dr. P. Thirunavukarasu received the received the B.Sc., M.Sc. and M.Phil degree in Mathematics from the Bharathidasan University, Tamilnadu, South India. He completed his Ph.D degree from Bharathidasan University/Regional Engineering College. He has published many papers in International and National level conferences. He also published many books. He is the Life member of ISTE and The Mathematics Teacher/JM/Books/official Journal of the Association of Mathematics Teachers of India. His research areas are Applications of Soft Computing, Analysis, Operations Research, Fuzzy Sets and Fixed point theory.

Mrs. A. Thanithamil received the received the B.Sc., M.Sc. degree in Mathematics from the Bharathidasan University, Tamilnadu, South India. She completed her M.Phil degree from Madurai Kamaraj University. She got B.Ed degree from Annamalai University and PGDCA from Bharathidasan University. She has published many papers in International and National level conferences. Her research areas are Applications of Graph theory, Analysis, and Algebra.