Int J Pharm Bio Sci 2013 Oct; 4(4): (B) 497 - 503

Research Article **Bioinformatics**

International Journal of Pharma and Bio Sciences **ISSN 0975-6299**

SIMULATED ANNEALING MODEL FOR RETICULATE EVOLUTION IN MOLECULAR SEQUENCES

ASHOK KUMAR DWIVEDI[*1] AND DR. USHA CHOUHAN

[1]Department of Mathematics, Bioinformatics and Computer Applications, Maulana Azad National Institute of Technology, Bhopal
[2]Department of Mathematics, Bioinformatics and Computer Applications, Maulana Azad National Institute of Technology, Bhopal

ABSTRACT

The molecular sequence in the nature are evolved rather than invented. They are adopted from preexisting sequences. The conventional assumption on of independent evolution has been challenged by the biological evidence of hybridization, horizontal gene transfer and recombination. In such a scenario classical approaches of representing evolution by phylogenetic tree construction method fails to represent evolution. To represent such an evolution we need network models called reticulate networks. Inference of such a network given some optimization criterion like maximum likelihood or maximum parsimony have been proved to be NP-complete, so we need a model which can provide a feasible low cost solution. In this paper we propose a model based on simulated annealing which provides a global optimum solution. This method is evaluated with other available method which gives promising results.

KEYWORDS: Phylogenetics, Reticulation, Reticulate Evolution, Simulated Annealing, Reticulate Networks.

ASHOK KUMAR DWIVEDI
Department of Mathematics, Bioinformatics and Computer Applications,
Maulana Azad National Institute of Technology, Bhopal

INTRODUCTION

Reticulation is the lack of independence between two evolutionary lineages. In this process two or more independent evolutionary lineages are combined at some level of biological organization[1].Reticulation can occur at different level of chromosomes, genomes and species. At the species level the reticulation event occurs when two lineages recombine to create a new one and horizontal gene transfer by which genes are transferred across species. At the population level, sexual recombination causes evolution to be reticulated, whereas meiotic recombination causes shuffling of genes at the chromosomal level. Reticulation events at the species level fail to be modeled by a bifurcating tree[2]. Traditional method for evolutionary analysis employs phylogenetic models. Phylogenetic models are based on the underlying assumption of independent evolution. But for much organism genetic exchange occurs between lineages to produce new independent lineages. Phylogenetic trees models are not much suitable to model such evolutionary concepts. These exchanges cannot be represented by tree like structures however they can be represented by network like structures such an evolution is called reticulate evolution and network structure representing them is called reticulated network. In such a scenario reticulate networks provides more suitable models for the evolutionary analysis. Reticulate network models can also be employed to study the organisms evolved as tree like models in such a case reticulate networks are used to resolve the conflicts in data set caused by incomplete lineage sorting or by inadequacies of assumed evolutionary model[3]. Reticulate network construction problem are computationally intensive and have been proved to be NP-hard[4].So we need approaches which can provide efficient solution for these problems. Sneath et al. were among the first who provided the methods for studying the mechanisms of reticulate evolution. Several other methods have been proposed for the identification of reticulate evolution in nucleotide sequences. They include tests for clustering[5], a randomization approach[6] Method developed by H. J. Benelt and A. Dress[7], is based on split decomposition which enables the representation of data in the form of a splits graph revealing the conflicting signals contained in the data. In a splits graph, a pair of nodes may be linked by a set of parallel edges depicting alternative evolutionary hypotheses. P. Lengendre and V. Markenkov[8] proposed to use reticulograms for detecting reticulation events in evolutionary data. They developed a distance-based method to infer reticulate phylogenies. That method uses the topology of a phylogenetic tree as a supporting structure for building a reticulogram. Another method MC-Net[9] present a heuristic algorithm to find an optimal circular ordering based on the Monte-Carlo method, called MC-Net algorithm. Circular ordering produced by the MC-Net is closer to optimal circular ordering than the N-Net[10] ,Furthermore, the networks corresponding to outputs of MC-Net made by SplitsTree are simpler than N-Net. A model for direct construction is proposed for reticulate network as a model for describing the evolutionary history of a set of sequences under recombination events using directed acyclic graphs(DAG's)[11] ,They also described a set of properties that a DAG must possess in order to provide a realistic model of recombination. M. T. Hallet and J. Lagergran[12] described a set of conditions on rooted DAGs to use them as models for evolution under lateral transfer events. Giudici and Castelo[13] proposed an elegant approach for moving from one DAG to another via three types of simple moves: adding a new edge, removing an existing one or reversing an existing edge. This method can be used for generating alternative topologies. Approaches which deal with the problem of reticulation in molecular sequence are very specific because of inherent computational complexity of constructing reticulate network. Despite of having several algorithms there is no generalized algorithm which constructs a reticulate networks directly given some optimization criteria. In this paper we propose a

method which of constructing maximum likelihood DAG using simulated annealing for representing the reticulate network.

MATERIALS AND METHODS

A reticulate network can be represented by a directed acyclic graph (DAG) $N = (V,E)$ which satisfies the following constraints[14]. The set V of nodes is partitioned into two sets. V_T : The set of tree nodes, a node $v \in V$ is a tree node if and only if one of these three conditions holds: (1) indgegree (v) = 0 and outdegree(v) =2 for root noded,(ii) indegree(v)=1 and outdegree(v) = 2 for leaf node or (iii) indegree(v) =1 and outdegree(v) =2 for internal node. V_N : The set of network nodes, a node $v \in V$ is a network node if and only if one of these two conditions holds- (i) indegree(v) =2 and outdegree(v) =1.(ii) indegree(v)=2 and outdegree(v) =1.clearly we have $V_T \cap V_N = \phi$ and can easily verify that we have $V_T \cup V_N = V$.The set of edges is partitioned into two sets:E_T : The set of tree edges: An edge $e = (u,v) \in E$ is a tree edge if and only if v is a tree node.E_N :The set of tree edges: An edge $e = (u,v) \in E$ is a network edge if and only if v is a network node.In such a network, a species appears as a directed path p that does not contain any network edge. If p_1 and p_2 are two directed paths that define two distinct species, then p_1 and p_2 must be edge disjoint, that is, the two paths cannot share edges. Let $\mathcal{G}$ be the set of directed acyclic graphs (DAGs) with vertices. Let $f: \mathcal{G} \longrightarrow \mathbb{R}$ be a real valued function over $\mathcal{G}$. The optimal graph is one which optimizes $f(G)$ $where$ $G \in \mathcal{G}$. We need to use soft computing approaches like simulated annealing to find an approximate global solution. Finding the reticulated network which best represent the given molecular sequence is a combinatorial optimization problem. It can be completely characterized by the search space S and the cost function or objective functions. In the maximization case the desired optimal solution x_{opt} is one for which

$$f(x_{opt}) \geq f(x) \; for \; all \; x \in S$$

The problem then can be stated simply as

$$Maximise(x) \; x \in S$$

The solution X_{opt} is called a global optimum and its objective value, the optimal cost is denoted by

$$f_{opt} = f(X_{opt}).$$

In order to find the global optimum solution simulated annealing technique is applied which was first introduced by Kirkpatrick et al[15]. The simulated annealing method is known to be a practical method for finding a global optimal solution for many combinatorial problems that otherwise requires an exhaustive search to find the optimal solution. Simulated annealing is inspired by physical annealing in which a solid is cooled very slowly, starting from a high temperature, in order to achieve a state of minimum internal energy. It is cooled slowly so that thermal equilibrium is achieved at each temperature. Thermal equilibrium can be characterized by the Boltzmann distribution[15].

$$P_T(X = x) = \frac{e^{-E_i/K_B T}}{\sum_{all\ states\ i} e^{-E_i/K_B T}} \qquad \text{(i)}$$

Where X is a random variable indicating the current state, E_X is the energy of state x, k_B is the Boltzmann's constant and T is temperature. The evolution of the state of a solid in a heat bath toward thermal equilibrium can be efficiently simulated by a simple algorithm based on Monte Carlo techniques which was proposed by Metropolis[15]. The Metropolis algorithm takes the current state x, and generates a new state y by applying some small perturbation. The transition from state x to state y is then accepted with probability

$$P_{accept}(x, y) = \begin{cases} 1, & if\ E_x - E_y \leq 0 \\ e^{-(E_x - E_y)/K_B T}, & if\ E_x - E_y > 0 \end{cases} \qquad \text{(ii)}$$

If accepted, y becomes the current state and the procedure is repeated. This acceptance rule is known as the Metropolis criterion. Here E_x *and* E_y are energies at state x and y respectively. Given a DAG G = (V, E) we call DAG G' = (V, E') a perturbation of G if and only if we can obtain G' from G with a single move by adding a new edge, removing or reversing an existing edge in G. If f is given objective function and two DAGs, G_1, G_2 we call d (G) = f (G_1) – f (G_2) is the difference of sum of the pair wise difference of optimization criteria taken over all the nodes in the directed acyclic graph. Now, given a DAG G = (V, E) and a random move proposed by the Simulated Annealing we need to:

1. Examine whether move is legal by examining whether or not move introduces a cycle.
2. Decide whether to accept the move based on the probability $e^{-d(G)/T}$
3. Perform the move.
4. The Algorithm-

One of the appealing features of the simulated annealing algorithm is its ease of implementation. The basic algorithm can be coded in only a few lines, as illustrated by

```
t = Initial tree;
f = Cost(x);
T = Initial Temperature ();
do {
do {
new x = Apply Perturbation To(t);
Δf = Cost (new t) - f;
if ( (Δf<0) OR (random[0,1]<exp(-Δf/T)) ) {
t = new t;
f = f + Δf;
}
} while Not At Equilibrium ();
T = Update Temperature (T);
} while Exit Condition Not Met ();
```

There are four basic components which must be specified for each particular optimization problem:

1. The search space.
2. The objective function or cost function to be minimized on the given search space.
3. The perturbation mechanism used to generate a new solution from the current one.
4. The *cooling schedule* which includes the initial temperature, the procedure for updating the temperature, and the termination criterion used to determine the stopping point for the algorithm. The search space and cost function are specified as part of the formulation of the problem, while the perturbation mechanism and cooling schedule deal more with the parameters which control the search

RESULTS

In this section we describe the procedure to obtain phylogenic networks from consensus tree generated by simulated annealing algorithm. Firstly phylogenetic trees are generated using above simulated annealing method then these trees are supplied to SplitsTres4 software for constructing phylogenetic network. We test the model on Adh-sequence data from 11 species[16]. A consensuses species tee has been derived from resulted trees using MEGA4[17]. shown in figure 1.Pylogenetic network shown in figure 2 is constructed using Split tree software[18].

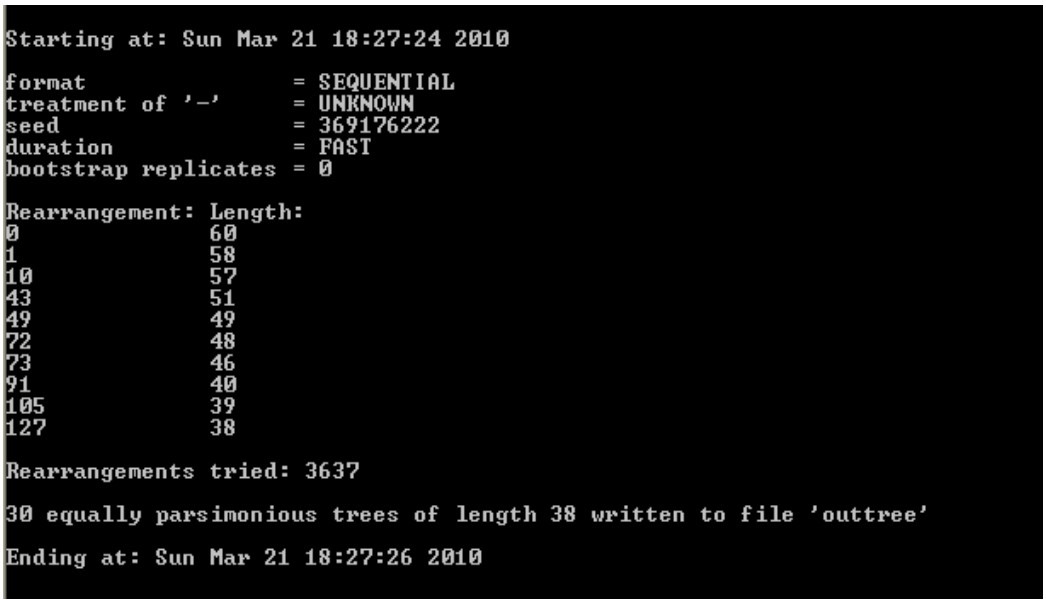

Figure 1
Output of Simulated Annealing in the search of most parsimonious trees using LVB

Table 1
Tree Pair Analysis for Reticulation event capturing using PhyloNet

Tree Pairs	Total Reticulations
T1/T2	3
T3/T4	2
T5/T6	3
T7/T8	3
T9/T10	6
T11/T12	2
T13/T14	2
T15/T16	1
T17/T18	4
T19/T20	4
T21/T22	3
T21/T22	2
T23/T24	1
T25/T26	4
T27/T28	2
T29/T30	3

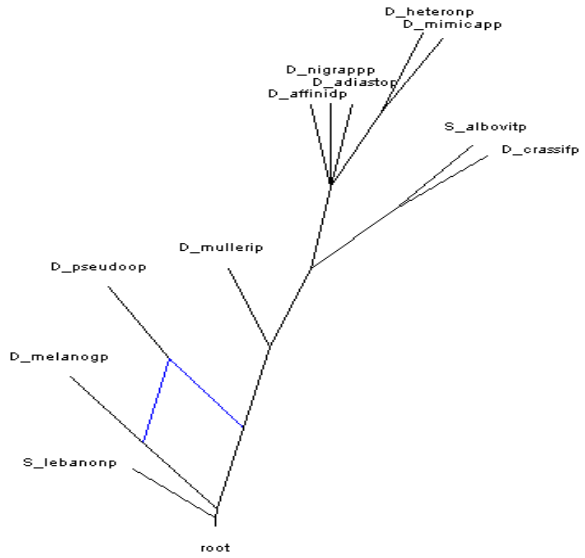

Figure 2
Phylogenetic network drawn by phylnet software

DISCUSSION

Phylogenetic trees generated using a simulated annealing algorithm tends to preserve biological signal of reticulated evolution as captured by in experiments. D_melanogp is displayed as result of reticulation events between S_labanonp and D_pseudoop as in figure 3.

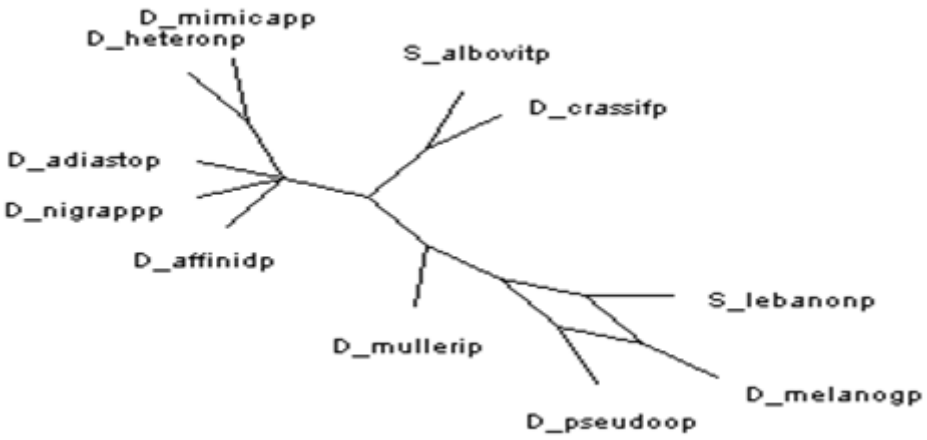

Figure 3
Network representation of tree using SplitTree 4 software

CONCLUSION

The heuristic search techniques as applied in various search and optimization problem can be applied in the field of Phylogenetics and particularly in reticulated evolution, this methods gives good results, in the sense of number of reticulation events captured, as it is a global search method.

ACKNOWLEDGEMENT

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. R. G.Beiko, W. F. Dollittle and and R. L. Charlabosis, The impact of reticulate evolution on genome phylogeny.Syst. Biol.,57:844-856,(2008).
2. H. Zhang, C. Y. Yu and B. Singer, Cell and tumour classification using gene expression data: construction of forests. Proc Natural Acad. Of Science, 4168-4172,(2008).
3. J. Hein, A heuristic method to reconstruct the history of sequences subject to recombination. Journal of Molecular Evolution,36(1993).
4. F. Murtagh, Complexities of Hierarchic Clustering Algorithms: the state of the art. Computational statistics quarterly,1,101-113,(1984).
5. J. C. Stephens, Statistical method for DNA sequence analysis: Detection of intragenic recombination or gene conversion, Molecular Biology Evolution,2,539-556,(1985).
6. S. Sayer, Statistical tests of detecting gene conversion, Mol Biol Evol,6,526,536(1989).
7. H. J. Bendelt and A. Dress, A canonical decomposition theory for matrices on a finite set, Advanced Mathematics, 92, 27-65,(1992).
8. P. Legendre and V. Makarenkov, Reconstruction of biogeographic and evolutionary networks using reticulograms, Syst. Biol.,51,199-216,(2002).
9. C Eslahchi, M. Habibi, R. Hassanzadeh, and E. Mottaghi, A method for the construction of phylogenetic networks based on Monte-Carlo method, BMC Evol Biol,10(2010).
10. D. Bryant and V. Moulton, Neighbor-Net: Aggalomerative method for the construction of phylogenetic networks.Mol Biol Evol.21,255-265(2004).
11. K. Strimmer and V. moulton, Likelihood analysis of phylogenetic networks using directed graphical methods, Molecular Biology and Evolution,17,875-881(2000).
12. M. T. Hallet and T. Lagergren, Efficient algorithm for lateral gene transfer problems.,In Proceeding of 5th Ann Int Conf Comp Mol Biol (RECOMB 01) New York .149-156(2001).
13. P. Giudici and R. Castelo,Improving marko chain monte carlo model search for data mining,Machine Learning,50,127-158,(2003).
14. Metropolis, Equations of state calculation by fast computing machines,Journal of chemical physics,21,1087-1092,(1953).
15. S Kirckpatrick, C. C. Gelatt and M. P. Vecchi, Optimization by simulated annealing, Science
16. R. H. Thomas and J. A. Hunt, Phylogenetic relationships in Drosophila: A confilict between molecular and morphological data.Mol Bio Evol,10,362-374,(1993).220, 671-680,(1983)
17. K. Tamura, Moleculr Evolutionary Genetics Analysis(MEGA) software version 4.0,Molecular Biology and Evolution.24,1596-1599,(2007)
18. D. H. Huson, SplitTree: analyzing and visualizing evolutionary data, Bioinformatics,14,68-73,(1998)