



International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

www.iasir.net

Genetic Algorithm Based Model for Reticulate Evolution in Molecular Sequences

Mr. Ashok K Dwivedi¹, Dr. Usha Chouhan²

^{1,2}Department of Mathematics, Bioinformatics and Computer Applications,
Maulana Azad National Institute of Technology (MANIT),
Bhopal, Madhya Pradesh, INDIA.

Abstract: Phylogenetic networks are models of the evolution of a set of organism that generalize phylogenetic trees. These models allow the existence of reticulation events such as recombination, hybridization, or horizontal gene transfer. These concepts are modeled as directed acyclic graphs rather than trees. We consider the problem of modeling such reticulate evolution using genetic algorithms. The algorithm has to be designed which incorporate various constrained by directed acyclic graph and the location of reticulate event. The choice of fitness function is also studied, and several possibilities for this purpose are presented and compared. The experimental evolution indicates that the genetic algorithm can satisfactorily recover the underlying evolution model behind the data. A computationally light fitness function seems to provide the best performance.

I. Introduction

Evolutionary history of collection of organism is represented by phylogenetic trees. This hierarchy shows the degree of closeness among the organism under study. As it turns out, inferring the best hierarchy is a formidably difficult task under several formulations (Day et al., 1986; Foulds and Graham, 1982). This hardness barrier can be circumvented using heuristic approaches. Genetic algorithm has been used in this domain with encouraging results (Lewis, 1998). But trees oversimplify our view of evolution, as it has been long recognized by biologists. There are many events in natural evolution in which the genetic material is not transferred in hierarchical way, e.g., hybrid speciation, horizontal gene transfer, etc. These phenomena, 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 (Huson and Bryant, 2006). No single methodology for the network construction is widely accepted to date (Moret et al., 2004b). For example, the detection and the identification of reticulation events had been approached by Halle et al. (Hallett et al., 2004) (focusing on horizontal gene transfer), and by Posada et al. (Posada and Crandall, 2002) (focusing on recombination events). Ankle et al. (Nakhleh, 2010) have proposed a method that combined preexisting consensus trees into a network with a single reticulation events. The methods mentioned above are in general based on deterministic approaches for finding good solutions, and hence the well known limitation arising from the $P \neq NP$ conjecture apply. To the best of our knowledge the inference problem has not been approached by metaheuristic techniques so far. However this later approach seems natural in this domain, given the success history of these techniques on the inference of phylogenetic trees. In this work, we propose an evolutionary approach to the phylogenetic network inference problem, and show that it can be a useful tool in this domain.

II. Methods

Some evolutionary events such as hybridization, recombination and horizontal genes transfer cannot be represented by phylogenetic tree model they are represented by directed acyclic graphs (DAGs). In the following section we will describe the notation used henceforth as well as some crucial notations such as time coexistence, and topological distance metrics on phylogenetic networks.

A. Notation

Let $G(V, E)$ be a DAG. We will use the notation $E(G)$ and $V(G)$ to denote respectively the set of edges and vertices of a DAG G . A directed path P of length k from u to v in a graph $G(u, v \in V(G))$, is a sequence $P = v < u_0, u_1, \dots, u_k >$ of nodes where $u = u_0$, $v = u_k$, and $(u_i, u_{i+1}) \in E(G)$ for $0 \leq i < k$. Let $\alpha(P) = u_0$, and $\omega(p) = u_k$ be the endpoints of path P . A node v is reachable from u in G if there exists a directed path from u to v ; in that case, u is an ancestor of v . Unlike trees, there may be more than one directed path between two nodes in a DAG. These paths are also termed as positive time directed paths for reason that will be clear at later point.

We can now define in-degree $\delta^\downarrow$ of a node as the numbers of edges arriving to that node and out-degree $\delta^\uparrow$ as the numbers of edges that depart from that node. There are some degree constraints on in DAGs representing phylogenetic networks. To be precise, a node $v \in V(E)$ is a tree node if (a) $\delta^\downarrow(v) = 0$

and $\delta^\uparrow(v) = 2$ [root (unique)], (b) $\delta^\downarrow(v) = 1$ and $\delta^\uparrow(v) = 0$ [leaf] or (c) $\delta^\downarrow(v) = 1$ and $\delta^\uparrow(v) = 2$ [internal tree node]. If a node is not a tree node, then it must have $\delta^\downarrow(v) = 2$ and $\delta^\uparrow(v) = 1$. Such nodes are termed as network nodes. An edge $e = (u, v) \in E(G)$ is a tree edge if and only if v is a tree node, and it would be a network edge otherwise. Notice that tree nodes describe mutations, and network nodes describe reticulation events.

B. Time Coexistence

A crucial consideration that must be taken into account in phylogenetic networks is the fact that each reticulation event defines a simultaneity plane: in order to have two species recombining their genomes, or having some genetic information transferred from a species to another, they must coexist in time. This way, the set of nodes $V(G)$ of g , a phylogenetic network G are implicitly ordered in the time according to the particular reticulation events existing in G . To be precise, we can assign a specific time stamp $t(v)$ to each node v in the network. Actually the absolute values of these time-stamps are not relevant; what matters is their relative ordering. Thus, if there exists a time directed path between node u and v , we would logically have $t(u) < t(v)$ i.e. u is an ancestor of v . However, if $e = (u, v)$ is a network edge, then $t(u) = t(v)$ because the reticulation events are instantaneous in the evolution time scale. Consider Fig. 1(b). Let $t(Y) = t_1$ and $t(X) = t_4$ and let reticulation events s_1 and s_2 happen at time t_2 and t_3 respectively. Now notice that there does not exist a positive time directed path from Y to X . However, even though Y is not an ancestor of X , it is impossible to have a reticulation event between these two nodes: Y is an ancestor of A that coexists with B ; B is in turn an ancestor of C that coexists with D , an ancestor of X . Hence $t_1 < t_2 < t_3 < t_4$. More formally, we say that two nodes $u, v \in V(G)$ cannot coexist in time if: U is an ancestor of v (or vice versa), or

There is a sequence of positive time directed paths $P = \{P_1, P_2, \dots, P_k\}$ such that $\alpha(P_1) = u$ (resp. v), $\omega(P_k) = v$ (resp. u), and for $1 \leq i < k$, there exists a network node whose parents are $\omega(P_i)$ and $\alpha(P_{i+1})$.

Time coexistence thus imposes constraints on where reticulation event can take place, and therefore on which DAGs actually represent a feasible phylogenetic network. These constraints will have to be observed when evolving networks within the inference algorithm.

C. Topological Distance Matrices on Phylogenetic Networks

Metrics for measuring the topological distance between networks are essential to interpret the result of an inference algorithm. They can be used to determine to which extent the features of a target network have been successfully recovered. For this purpose, we can resort to generalization of well known distance matrices for these trees, such as the Robinson-Foulds (RF) distance (Robinson and Foulds, 1981). The RF distance on trees uses the notation of bipartition: given an edge in a tree T , we can partition the leaf set $\mathcal{L}$ of T into the disjoint sets $A(e)$ and $C(e)$, respectively comprising the leaves in that is reachable from the root (resp. unreachable) through edge e . The notion of bipartition in trees is readily generalizable to tripartitions in network. In this latter case, and edge e induces tripartition $\langle A(e), B(e), C(e) \rangle$, where $A(e)$ comprising the leaves that are reachable from the root only through edge e , $B(e)$ comprises the leaves that are reachable from the root by a path goes through e and at least by another path that does not pass through e , and $C(e)$ is defined as before.

We denote by $\square(e) = \langle A(e), B(e), C(e) \rangle$ the tripartition induced by the e . Two tripartition $\square(e_1)$ and $\square(e_2)$ are equivalent $\square(e_1) \equiv \square(e_2)$ if, and only if, $A(e_1) = A(e_2)$ and $B(e_1) = B(e_2)$ and $C(e_1) = C(e_2)$. Now, two edges e_1 and e_2 are compatible ($e_1 \equiv e_2$) if and only if $\square(e_1) \equiv \square(e_2)$. Let $\delta: \mathbb{R} \rightarrow \{0, 1\}$ be defined as $\delta(\text{TRUE}) = 1$ and $\delta(\text{FALSE}) = 0$. Let $\Gamma(G_1, G_2)$ be defined as

$$\Gamma(G_1, G_2) = \frac{1}{|E(G_1)|} \sum_{e_1 \in E(G_1)} 1 - \sum_{e_2 \in E(G_2)} \delta(e_1 \equiv e_2) \quad (1)$$

It is then possible to define the false negative rate $FN(G, \tilde{G}) = \Gamma(\tilde{G}, G)$, and the false positive rate $FP(G, \tilde{G}) = \Gamma(G, \tilde{G})$ between the inferred network G and the target network $\tilde{G}$. Finally, the RF distance for networks can be estimated as $D_{RF}(G, \tilde{G}) = (FN(G, \tilde{G}) + FP(G, \tilde{G}))/2$. Notice that the RF distance is 0.0 for two identical network, and 1.0 for two networks without any compatible edge.

III. Genetic Algorithm for Inferring Phylogenetic Network

In order to tackle the inference of phylogenetic networks with evolutionary algorithm, we consider a direct approach in which the search is directly conducted on the space of all possible phylogenetic networks with given leaf set. Thus each individual in the population of the EA represents a feasible phylogenetic network, internally encoded as an adjacency matrix. This means that (i) an initialization process producing feasible networks must be used, and (ii) the reproductive operators used must respect feasibility, i.e., they must always produce feasible offspring. The details of these operators will be described in section A.

Another central element in this EA is the fitness function. The RF metric defined in the previous section can be used for the off-line assessment of the results, but cannot obviously be used during the evolution. On the contrary, the fitness function must evaluate a phylogenetic network on the basis of the particular dataset of

molecular sequences to be modeled. Several choices have been explored for this purpose. These are described in section B.

A. Evolutionary Operators

The first issue to be tackled in the genetic algorithm is the generation of feasible networks for insertion in the initial population. To do so, each time a new network is required a random tree is firstly generated (this is done by firstly constructing a random permutation of leaves; then, an initial tree is built with the first two leaves in the permutation, and remaining leaves are subsequently inserted at random points of the partial tree until it is finally completed). Once the tree has been obtained, network nodes are inserted by randomly selecting pairs of tree nodes that can coexist in time. After having generated a population of feasible networks, adequate reproductive operators must be used. Let us firstly consider the recombination operator. As usual, this operator takes information pieces from two individuals, and combines them to create a new feasible solution. In this case, these information pieces take a shape of sub networks, and hence we can express the process in terms pruning and grafting sub networks. Let G_1 and G_2 be the networks to recombine, and let trees be represented in LISP notation. The process is follows:

Select a random sub network N (rooted at a tree node) from

For each leaf $u \in N$ do

Find sub network U in G_2 such that $U = (h, (u), U')$ or $U = (h, U', (u))$.

Replace U by U' in G_2 .

Select a random sub network V from G_2 .

Replace V by $V' = (h', N, V)$, where h' is a new internal tree node.

This operator can be regarded as a generalization of the Prune-Delete-Graft (PDG) operator for trees [5, 6]. Thus we have named it NetPDG. Notice that some network node might be broken during the recombination process. This could happen either in step 1 (if there were a reticulation event between a node in N and another node not in N), or in step 2b (if a grandson of a network node were removed). As to the mutation operator, it is based on rearranging the topology of a portion of the network. More precisely to mutate network G , this operator selects a random sub network N from G , and generates another random network, spanning the same set of leaves, and having the same number of internal network nodes (network nodes with one parent in N and other parent outside N are broken). Notice that not only the number of network nodes in the child might be lower than that of the parental solution(s) (if some such nodes are broken during recombination- as described before or during sub network scrambling in mutation); it can be also higher than it if no network node is broken, and new ones are transferred during the recombination. In this work, we have opted for keeping a fixed, predefined number of network nodes in solution. Hence, whenever a new solution has a higher or lower number of network nodes, it is repaired (breaking randomly selected network nodes, or adding new ones).

B. The Fitness Function

As it is the case for phylogenetic trees, the accuracy with which a phylogenetic network model the evolutionary history of a certain dataset can be computed via sequence base methods (i.e. maximum parsimony (Nakhleh et al., 2005), or maximum likelihood (Strimmer and Moulton, 2000) and distance based methods (Makarenkov et al., 2005). Among these approaches, maximum likelihood is an appealing way of assessing the quality of a proposed phylogenetic model: they consider all possible evolutionary pathways compatible with the molecular data available, and are known to be asymptotically accurate (Huelsenbeck and Crandall, 1997). We have thus opted for maximum likelihood approach here. The general setting is the following: we have a collection D of n sequences, representing some molecular data from n different species. Here these sequences representing some molecular data from n different species. Here, these sequences are taken from the alphabet $\Sigma = \{A, C, T, G\}$ i.e. they represent DNA sequences. We assume a certain evolution model at the molecular level, indicating the likelihood that a certain character (nucleotide in this case) mutates into another one, say $\Pi = \{\Pi_{ij}\}_{i,j \in \Sigma}$. Each site in the sequence is assumed to evolve independently. Likewise, we assume a certain mechanics for reticulation events, i.e. network nodes indicate recombination, and these produce organisms whose genetic sequence is e.g., the result of a uniform crossover of the parental sequences. When this general evolutionary framework is superimposed on a particular network N , we can calculate $P(D|N)$, as described below.

The first method is based on the formula used for likelihood calculation in trees. Let L_{k,s_k}^r be the likelihood of a network rooted at node k , given that node has nucleotide s_k in site r . If node k is the parent of nodes i and j in the network, and both are tree nodes, then

$$L_{k,s_k}^r = \left(\sum_{s_i \in \Sigma} \Pi_{s_k,s_i} L_{i,s_i}^r \right) \left(\sum_{s_j \in \Sigma} \Pi_{s_k,s_j} L_{j,s_j}^r \right) \quad (2)$$

In case node k were a network node, the first term in Eq. (2) would have to be changed accordingly. To be precise, the state of node i would depend on the state of node k , and also on the state of the other parent, say node z . This bivariate dependency precludes the fast recursive evaluation of Eq. (2). To circumvent this issue, we can take into account the fact that, due to the semantics of recombination, the state of node k propagates with $\frac{1}{2}$

probabilities to node i . Thus, we approximate this first term as $\frac{1}{2} \cdot L_{v,s_k}^r$, where v is the unique child of node i in the network. The same reasoning would apply to node j were it a network node. While this is a mere approximation of the exact likelihood of these network nodes, it allows a very fast recursive evolution of the overall likelihood of the complete network. This evaluation is completed by noting that (i) the likelihood $L_{w,s}^r$ of a leaf w is 1.0 if the r^{th} site of w^{th} sequence is s , and 0 otherwise, and (ii) the complete likelihood of the network for site r is $L^r = \sum_{s_0 \in \Sigma} \prod_{s_0} L_{s_0}^r$, where $\prod_{s_0}$ is the marginal probability of nucleotide s_0 . Finally, since sites evolve independently, the likelihood of the network for the whole data is $L = \prod_{i=1}^m L^i$, m being the length of sequences. We term this evolution method ABE (after approximation Bayesian estimation).

Monte Carlo (MC) methods constitute an alternative to ABE providing an asymptotically exact numerical estimation of the network likelihood. This estimation is obtained by constructing N samples of the states of internal nodes in the network, and computing

$$L^r = \frac{1}{N} \sum_{i=1}^N \prod_{w'} \Pi_{s_w^i, s_w} \quad (3)$$

Where the inner product ranges over all leaves w of the network, w' is the parent node of a certain w , s_w^i is the i^{th} sample state in the r^{th} site for node w' , and s_w is the actual in the r^{th} site of w^{th} sequence. In order to have a correct MC integration, the probability of each sample must be a proportional to its real likelihood. This can be easily accomplished by assuming a random state at the network root (following the marginal probabilities $\prod_{s_0}$) and simulating the evolution of this site along the network, using the stochastic model is considered.

The MC method provides an asymptotically more accurate estimation of the exact likelihood, but it has a much higher computational cost than ABE. In order to alleviate this cost partially combined methods (CMB) has been considered. The basic idea is identifying all the sub trees in the network (that is, maximal sub graphs without network nodes), using the MC method just to sample the state for remaining nodes. Subsequently,

$$L^r = \frac{1}{N} \sum_{i=1}^N \prod_v L_{v,s_v^i}^r \prod_w \prod_{w'} \Pi_{s_w^i, s_w} \quad (4)$$

Where the first product ranges over all internal nodes v being the root of a maximal sub tree, s_v^i is the corresponding value in the i^{th} sample, the second product ranges over all leaves that are not part of a maximal sub tree, and w' , s_w^i , and s_w are interpreted as before. Thus, the exact likelihood value is computed for maximal sub trees, and the cost of the MC component is reduced.

IV. Experimental Results

The data used in the experiments have been synthetically generated from known evolutionary models, in order to allow an objective measurement of the extent to which the inference algorithms were capable of recovering the underlying model. The process consist of generating a random network with the desired number of leaves (n) and network nodes (k), and then constructing nucleotide sequences by simulating the evolution of r sites throughout the network. The stochastic evolution of sequences is done assuming the Kimura 2-parameter model (Kimura, 1981) with transition rate $\alpha = 0.05$ and transversion rate $\beta = 0.025$. We have considered network with $n \in \{10, 25\}$ leaves, $k \in \{0, 1, 2\}$ network nodes, and $r \in \{100, 250\}$ sites per sequence. Both the procedure for generating the dataset, and the parameters used are similar in related works (Nakhleh et al., 2004). A steady-state evolutionary algorithm with standard parameters ($popsize = 100$, $px = .9$, $p_m = 1/l$, l being the number of nodes, $maxevals = 1000n$, binary tournament selection) has been used in experiments. No fine tuning of these parameters has been attempted. To allow a wider exploration of the capabilities of the evolutionary algorithms, a different problem instance has been used in each run. This way we are evaluating the algorithm on many samples of the whole problem class, rather than just on a couple of instances. Results have been obtained for the three fitness function described earlier. Twenty runs have been done for each parameter set of functions MC and CMB. Function ABE turns out to be around 50 times faster than MC (using 500 samples per evolution), so we have conducted 1000 runs for it per parameter set. The best networks found are evaluated in term of RF distance with respect to the real network. For the network model considered, the most related approach in the literature is (Nakhleh et al., 2004).

V. References

- [1]. Day, W., D, J., D, S., 1986. The computational complexity of inferring rooted phylogeny by parsimony. *Mathematical biosciences* 81, 33-42.
- [2]. Foulds, L., Graham, R., 1982. The Steiner Problem in phylogeny is NP-complete. *Advances in Applied Mathematics* 3, 439-449
- [3]. Hallett, M., Lagergren, J., Tofigh, A., 2004. Simultaneous identification of duplications and lateral transfers. 8th Annual International Conference on Computational Molecular Biology, pp. 347-356.
- [4]. Huelsenbeck, J., Crandall, K., 1997. Phylogeny estimation and hypothesis testing using maximum likelihood. *Annual Review of Ecology and Systematics* 28, 437-466.

- [5]. Huson, D.H., Bryant, D., 2006. Application of phylogenetic networks in evolutionary studies. *Molecular biology and evolution* 23, 254-267.
- [6]. Kimura, M., 1981. Estimation of evolutionary distances between homologous nucleotide sequence. *Natural Academy of Sciences*, pp. 437-446.
- [7]. Lewis, P., 1998. A genetic algorithm for maximum-likelihood phylogeny inference using nucleotide sequence data. *Molecular Biology and Evolution* 15.
- [8]. Makarenkov, V., Kevorkov, D., Legendre, P., 2005. Modeling phylogenetic relationships using reticulated networks. *Zoologica Scripta* 33, 89-96.
- [9]. Moret, B., Nakhleh, L., Warnow, T., Linder, C., Tholse, A., Padolina, A., Sun, J., Timme, R., 2004a. Phylogenetic networks: Modeling, reconstructability, and accuracy
- [10]. *IEEE transactions on Computational Biology and Bioinformatics* 1, 13-23.
- [11]. Moret, B., Nakhleh, L., Warnow, T., Linder, C., Tholse, A., Pasdolina, A., Sun, J., Timme, R., 2004b. Phylogenetics networks: Modeling, reconstructability. and accuracy. *IEEE transactions on Computational Biology and Bioinformatics* 1, 13-23.
- [12]. Nakhleh, L., 2010. Evolutionary Phylogenetic Networks: Models and Issues. 125-158.
- [13]. Nakhleh, L., Jin, G., Zhao, F., Meller-Crummey, J., 2005. Reconstructing phylogenetic networks using maximum parsimony. *Computational Systems Bioinformatics Conference. IEEE press*, pp. 93-102.
- [14]. Nakhleh, L., Warnow, T., Linder, C., 2004. Reconstructing reticulate evolution in species- theory and practices. *Annul International Conference on Computational Molecular Biology*, pp. 337-346.
- [15]. Posada, D., Crandall, K.A., 2002. The effect of recombination on the accuracy of phylogeny estimation. *Journal of molecular evolution* 54, 396-402.
- [16]. Robinson, D., Foulds, L., 1981. Comparison of phylogenetic trees. *Mathematical Bio-sciences*, 131-147.
- [17]. Strimmer, K., Moulton, V., 2000. Likelihood analysis of phylogenetic networks using directed graphical methods. *Molecular Biology and Evolution* 17, 875-881.

VI. Acknowledgements

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.