Phylogenetic and Phylogenomic Analyses for Large Datasets

Le Sy Vinh

University of Engineering and Technology, Vietnam National University, Hanoi, Vietnam

Email: vinhls@vnu.edu.vn

Communication: received 28 October 2019, revised 30 December 2019, accepted 30 December 2019

Digital Object Identifier: 10.32913/mic-ict-research.v2019.n2.898

The Editor coordinating the review of this article and deciding to accept it was Prof. Le Duc Hau

Abstract: The phylogenetic tree is a main tool to study the evolutionary relationships among species. Computational methods for building phylogenetic trees from gene/protein sequences have been developed for decades and come of age. Efficient approaches, including distance-based methods, maximum likelihood methods, or classical maximum parsimony methods, are now able to analyze datasets with thousands of sequences. The advanced sequencing technologies have resulted in a huge amount of data including whole genomes. A number of methods have been proposed to analyze the whole-genome datasets, however, numerous challenges need to be addressed and solved to translate phylogenomic inferences into practices. In this paper, we will analyze widely-used methods to construct large phylogenetic trees, and available methods to build phylogenomic trees from whole-genome datasets. We will also give recommendations for best practices when performing phylogenetic and phylogenomic analyses. The paper will enable researchers to comprehend the state-of-the-art methods and available software to efficiently study the evolutionary relationships among species from large datasets.

Keywords: *DNA sequence, evolutionary relationship, senome, phylogenetics, phylogenomics, large dataset, protein sequence.*

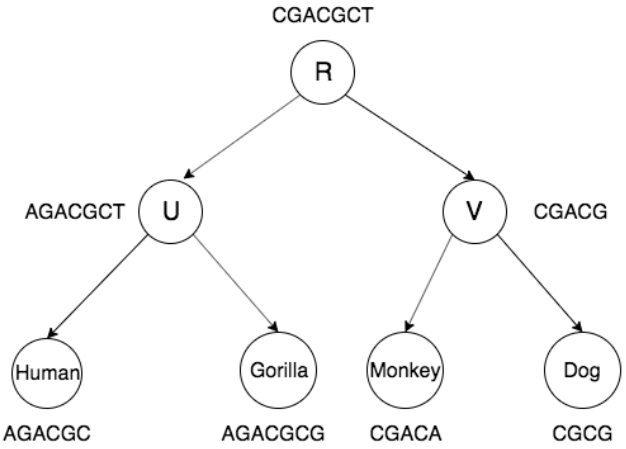

Figure 1. A phylogenetic tree represents the evolutionary relationships among species.

I. Introduction

Phylogenetic reconstruction is one of the most essential means to study the relationships among species. Analyzing the relationships among species shed light on the evolution of species. The phylogenetic information also enables us to study the structures and functions of DNA/protein sequences. Phylogenetic inference is an active research field for decades, and phylogenetic tree reconstruction methods based on molecular data have been developed and employed to study the evolution of species in tens of thousands of studies.

The evolutionary relationships among species are typically described by a binary tree where external nodes represent for present species; internal nodes represent for ancestors of species, and branches reflex connections and genetic changes between species (see Figure 1). Computational approaches have been proposed and implemented to construct phylogenetic trees based on several types of data, including morphological characters and molecular data, *i.e.*, DNA sequences, protein sequences, or even whole genomes. In this paper, we focus on phylogenetic inferences from molecular data.

The first class of phylogenetic tree construction approaches is distance-based methods. The phylogenetic tree is built based on genetic distances between sequences, *i.e.*, species with small distance should be placed closely in the tree. A number of distance-based methods have been proposed to reconstruct large phylogenetic trees [1–3]. Among them, the Neighbor-joining method and its improved modification [1, 2] have been used to build phylogenetic trees for tens of thousands of evolutionary studies.

The maximum parsimony approach searches the best phylogenetic tree that minimizes the number of changes to explain the difference among species. The maximum parsimony methods are simple, intuitive, and widely used to build phylogenetic trees [4–7]. However, they suffer a systematic limitation that they cannot describe complex changes occurred along branches of the tree. Thus, the maximum parsimony methods are suitable for analyzing datasets of closely related species.

The maximum likelihood approach has been developed to search for the tree that maximizes the probability of data [8–11]. The maximum likelihood methods are normally better than distanced-based and maximum parsimony methods, however, they are computationally expensive for large datasets. Heuristic approaches have been proposed to search maximum likelihood trees for datasets with thousands of species. Currently, maximum likelihood methods are most widely used for phylogenetic inferences.

The current sequencing technologies enable us to obtain whole genomes of thousands of species. Building phylogenomic trees based on whole-genome data becomes a common practice. Numerous challenges must be considered when analyzing whole genomes, including computational burden, the heterogeneity of evolutionary models for different genes, or gene structural variation in the genomes. Both computational methods, evolutionary models, and efficient software must be expanded to handle all the requirements when analyzing large genome datasets.

II. Data and Models

1. Data

Nowadays, molecular data, *i.e.*, DNA and protein sequences are the most common data type used to study evolutionary relationships among species. Each species is represented by one or multiple sequences, even by its whole genome. A DNA sequence (or genome) is coded as a string of 4 different nucleotides, *i.e.*, A, C, G, T; while a protein sequence is described as a string of 20 different amino acid types. Each nucleotide/amino acid in a DNA/protein sequence is called a character. Two characters in two genomes are called homologous if they have derived from the same common ancestor character. Generally, two DNA/protein sequences are called homologous if they originated from the same ancestor sequence. Note that we are only able to obtain molecular sequences from current species. In other words, we are not able to collect molecular data from common ancestor species for our studies. The data from ancestor species are normally inferred from the data of their descendants.

Three common kinds of character changes during the evolution are substitutions, insertions, and deletions. The substitutions will change the content of sequences. As a result, two homologous sequences originated from the same ancestor might be different. The insertion/deletion events (indels for short) during the evolution resulted in homologous sequences with different lengths. The first task in phylogenetic analysis is to align homologous sequences to create a multiple sequence alignment such that characters at the same column are assumed to be homologous (see Table I). The difference between homologous characters is phylogenetic signals reflexing the evolutionary relationships among species. A number of alignment methods have been proposed to align DNA/protein sequences, notably ClustalW [12] and Muscle [13]. They apply a progressive aligning strategy to build a multiple sequence alignment that minimizes the number of changes among sequences. Multiple sequence alignments are the input for phylogenetic or phylogenomic inferences.

TABLE I
A multiple sequence alignment of four sequences where '-' character represents for indels

	1	2	3	4	5	6	7
Human	A	G	A	C	G	C	-
Gorilla	A	G	A	C	G	C	G
Monkey	C	G	A	C	A	-	-
Dog	C	G	-	C	G	-	-

2. Models

The substitution process of characters during the evolution is complex. It can be typically simplified and modeled by a Markov process with following assumptions [14]:

- The rate of change from a character i to another nucleotide j is independent of the history of nucleotide i (Markov property);
- The substitution rate is time-homogeneous, *i.e.*, constant over the time course;
- The substitution between characters is time-continuous, *i.e.*, occurring at any time during the evolution process;
- The frequencies of characters are at equilibrium (stationarity).

The substitution process is represented by an instantaneous substitution rate matrix $Q = \{Q_{ij}\}$ where Q_{ij} is the number of substitutions from character i to character j per time unit. Note that matrix Q is a 4×4 matrix for nucleotide model or 20×20 matrix for amino acid model.

As usual, the substitution process in phylogenetic analyses is also assumed to be time-reversible, *i.e.*, the relative substitution rates between nucleotide i and nucleotide j are

TABLE II
THE SUBSTITUTION MODEL FOR NUCLEOTIDES

$Q =$		A	C	G	T
	A	$-r_{AC}\pi_C$ $-r_{AG}\pi_G$ $-r_{AT}\pi_T$	$r_{AC}\pi_C$	$r_{AG}\pi_G$	$r_{AT}\pi_T$
	C	$r_{CA}\pi_A$	$-r_{CA}\pi_A$ $-r_{CG}\pi_G$ $-r_{CT}\pi_T$	$r_{CG}\pi_G$	$r_{CT}\pi_T$
	G	$r_{GA}\pi_A$	$r_{GC}\pi_C$	$-r_{GA}\pi_A$ $-r_{GC}\pi_C$ $-r_{GT}\pi_T$	$r_{GT}\pi_T$
	T	$r_{TA}\pi_A$	$r_{TC}\pi_C$	$r_{TG}\pi_G$	$-r_{TA}\pi_A$ $-r_{TC}\pi_C$ $-r_{TG}\pi_G$

the same in both directions. As a result, the instantaneous substitution rate matrix Q can be decomposed into a symmetric relative substitution rate matrix $R = \{r_{ij}\}$ and character frequency vector π. Technically, $Q_{ij} = \pi_j r_{ij}$ if $i \neq j$, otherwise, $Q_{ij} = -\sum_{x \neq i} Q_{ix}$ (see Table II).

It is well known that the substitution rates are heterogeneous among character sites, *i.e.*, the substitution rates are normally fast at unimportant functional sites and slow at important functional sites. The rate models have been proposed to describe the rate heterogeneity. The widely-used rate model in phylogenetic analysis is the combination of a gamma rate model G and an invariant site rate model I [15]. The gamma rate model G assumes that the substitution rates at sites follow a gamma distribution that can be approximated by a discrete gamma distribution with several categories. The invariant rate model I assumes that a certain portion of sites in the alignment is invariant.

The parameters of substitution models and/or rate models can be directly estimated from the dataset under the study, except the amino acid substitution models. The amino acid substitution models consist of more than 200 parameters that cannot be properly estimated from small or medium datasets under the study. Normally, the parameters of amino acid substitution models are empirically estimated from large datasets [16–19].

III. PHYLOGENETIC ANALYSIS

Given $D = \{d_1, d_2, \ldots, d_n\}$ is a multiple sequence alignment of n sequences with length l, where n sequences representing n species, the phylogenetic tree reconstruction method will build a binary tree T with n leaves representing n sequences, and internal nodes representing ancestors. As we do not have data from ancestors, it is hard to determine the root of a phylogenetic tree. Therefore, the phylogenetic tree is typically represented by a binary unrooted tree. The number of binary unrooted tree structures with n leaves is $\prod_{i=3}^{n}(2i-5)$ that increases exponentially with the number of leaves. Searching the best tree for large n is computationally expensive. A number of heuristic approaches have been proposed to construct phylogenetic trees based on different criteria. In this section, we will discuss three widely-used approaches, including distance-based approach, maximum parsimony approach, and maximum likelihood approach, to build large phylogenetic trees.

1. Distance-based Approach

Analyzing the evolutionary relationships among species based on genetic distances among species have been introduced for more than fifty years [20]. The distance-based approach builds a phylogenetic tree such that closely related sequences with small distances should be placed nearby on the tree. The distance-based method comprises two steps: estimating the genetic distance matrix between sequences and constructing a tree based on the distance matrix. The genetic distance between two sequences can be efficiently estimated using the maximum likelihood method. We note that the accuracy of the estimated distance between two sequences increases with the length of sequences.

Let $D = \{d_{ij}\}$ be a distance matrix where d_{ij} is the genetic distance estimated between two sequences i and j. Let p_{ij} be the length of the path connecting two sequences i and j on the tree. The distance-based approach builds a tree such that the path length p_{ij} on the tree reflexes the distance d_{ij} on the distance matrix D for all pairs of sequences. Minimizing the total square difference is the best statistically justified distance-based objective [14]. Technically, the distance discrepancy $\Delta(T) = \sum_{u=1}^{n}\sum_{v=1}^{n}(d_{uv} - p_{uv})^2$ is the objective function of the least square method. The least square method will construct a tree T to minimize the distance discrepancy $\Delta(T)$. Given a tree structure, the first task of the least square method is the estimation of branch lengths to minimize the function $\Delta(T)$. This task can be efficiently solved by algebraic analysis [21]. As the number of tree structures increases exponentially, searching the least square tree is an NP-complete problem [22]. A number of heuristic distance-based methods have been proposed to solve the problem.

Among distance-based methods, neighbor joining method (NJ) or its modified version BioNJ is the most popular and widely used to construct large phylogenetic trees [1, 23]. The NJ algorithm starts from a star tree of n leaves, *i.e.*, each leaf is directly connected to the root and considered as a subtree. The NJ algorithm iteratively joins subtrees to build a whole binary unrooted tree (see

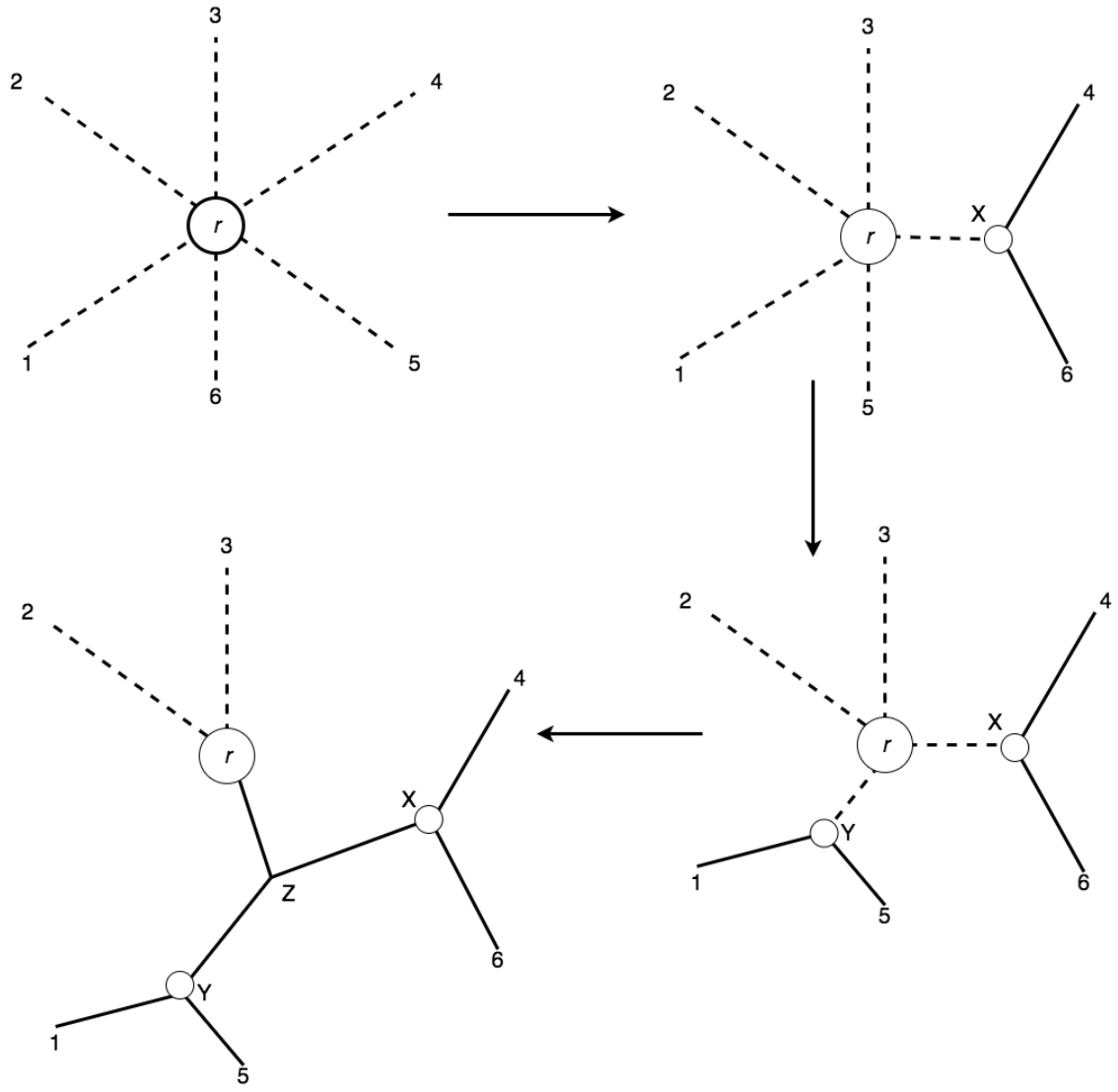

Figure 2. The neighbor joining algorithm to build a phylogenetic tree based on the distance matrix between sequences.

Figure 2). The total distance discrepancy based on the star tree is considerably large, therefore, the NJ algorithm step-by-step joins subtrees together to create a new subtree consisting of the two subtrees. The NJ algorithm iteratively joins two subtrees that helps to reduce the distance discrepancy at most. The complexity of the NJ algorithm is $O(n^3)$, thus, it is suitable to build distance-based trees for datasets containing several hundred sequences.

Other faster distance-based methods have been developed to build trees with larger datasets. Vinh and colleagues have proposed the shortest triplet clustering (STC) algorithm with the complexity of $O(n^2)$ to build distance-based phylogenetic trees for thousands of sequences (Vinh and von Haeseler 2005 [3]). Consider two leaves x and y, let r_{xy} be the most common ancestor of x and y. The key idea of the STC algorithm is the observation that if r_{xy} is the farthest internal node to the root of the tree, two leaves x and y should be neighbors on the tree. The STC algorithm uses the condition to iteratively group sequences to build a whole tree. Experiments on a wide range of simulated datasets showed that the STC algorithm was faster and more accurate than the NJ method. In practice, the NJ algorithm is still the most trusted and widely used distance-based method.

2. Maximum Parsimony Approach

Maximum parsimony approach is a simple and intuitive approach to construct phylogenetic trees from alignments. The approach searches for the tree that requires minimum number of changes along the tree (called the tree length) to explain the difference between sequences [24]. Given a tree with sequences at leaves, determining sequences at internal

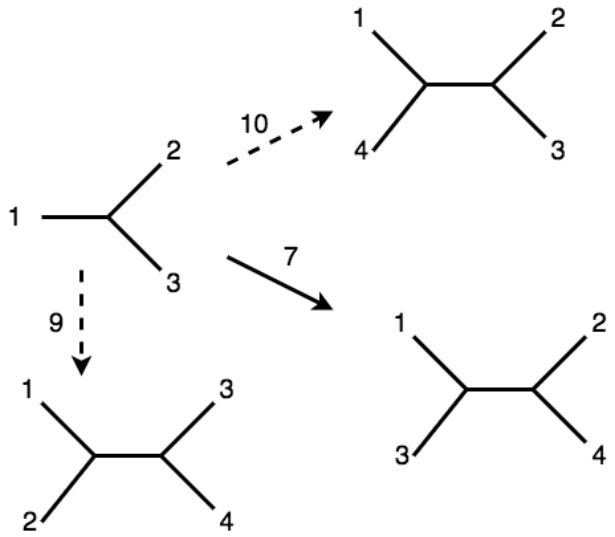

Figure 3. The stepwise addition algorithm to build a phylogenetic tree.

nodes to minimize the tree length can be efficiently solved by dynamic programming algorithm [25]. However, the number of tree structures increases exponentially with the number of sequences, searching the maximum parsimony tree is an NP-complete problem [26].

Several heuristic methods have been proposed to search the maximum parsimony trees. The key search strategy combines a stepwise addition tree construction method and a hill-climbing optimization. The stepwise addition tree construction method starts from a unique tree of three sequences, then iteratively adds new sequences into the current tree to construct a full tree of all sequences. A new sequence will be added into a branch of the current tree such that the length of the new tree is minimal (see Figure 3). As the stepwise addition tree construction method is a heuristic method, the constructed tree is normally still far away from the best tree. The constructed tree is further optimized by a hill-climbing method based on subtree rearrangements such as nearest neighbor interchange (NNI), subtree pruning and regrafting (SPR), or tree bisection and reconnection (TBR).

The combination of stepwise addition tree construction algorithm and hill-climbing optimization results in reasonable trees, however, they are still local optimal trees. The tree constructed by the stepwise addition tree construction algorithm depends on the order of sequences adding to the tree, *i.e.*, different sequence orders result in different trees. In an attempt to find the best maximum parsimony tree, we can perform the search process several times. The tree with the minimum length found will be considered as the best tree. Several software has been developed to build maximum parsimony trees from DNA/protein sequences such as PAUP* [4], TNT [5], or MPBoot [6]. The TNT and MPBoot have been designed to build maximum parsimony trees with large datasets including thousands of sequences. Note that, there might exist multiple most parsimonious phylogenies for the same set of sequences. To solve the problem, TNT and MPBoot software are also able to quickly build bootstrap trees to assess the reliability of constructed trees.

The maximum parsimony methods do not assume any substitution model to represent the substitution process of characters. As a result, the main drawback of maximum parsimony methods is the inability of reflexing complex changes of characters along branches of the tree, *i.e.*, branches are not able to present parallel substitutions, back substitutions, or multiple substitutions. Thus, the length of maximum parsimony tree might underestimate the true number of character changes when analyzing datasets with largely diverse sequences. In other words, the maximum parsimony methods are suitable for analyzing datasets consisting of closely related species.

3. Maximum Likelihood Approach

Maximum likelihood approach has been developed to overcome the limitation of maximum parsimony approach. Given a multiple sequence alignment $A = \{a_1, a_2, \ldots, a_n\}$ of n sequences with length l; and a substitution model M, the maximum likelihood (ML) method determines an unrooted binary tree T to maximize the probability of alignment A based on tree T and model M. Specifically, determining T and M such that the likelihood value $L(M,T;A) = P(A|M,T)$ is maximum where $P(A|M,T)$ is the conditional probability of alignment A given tree T and model M.

As usual, we assume that the substitution processes among sites are independent, and follow the same tree T and model M. Let a^j be nucleotides at position j^{th} in the alignment A, the likelihood value $L(M,T;A)$ can be calculated from the likelihood values of all sites as follows:

$$L(M,T;A) = \prod_{j=1}^{l} L(M,T;a^j), \tag{1}$$

where $L(M,T;a^j) = P(a^j|M,T)$, the conditional probability of alignment a^j given tree T and model M.

The rate heterogeneity among sites should be taken into account when calculating the likelihood of a tree T. We recall that the site rates can be described by a rate model V combining an invariant rate model and a discrete gamma distribution with C classes whose rates are $r_1, r_2, \ldots, r_C$,

respectively [27, Yang (1994)]. Technically, the likelihood value of $L(M, V, T; A)$ is calculated as follows:

$$L(M, V, T; A) = \delta \prod_{j=1}^{l} L(inv; a^j) + (1 - \delta) \prod_{j=1}^{l} \frac{1}{C} \sum_{c} L(M, r_c T; a^j), \quad (2)$$

where $L(inv; a^j) = P(a^j | inv)$, the conditional probability of a^j given that all characters at the site are identical (invariant), and $L(M, r_c T; a^j) = P(a^j | M, r_c T)$, the conditional probability of a^j where $r_c T$ is the tree T whose lengths are multiplied with rate r_c.

Given a tree structure with sequences at leaves, a critical task is determining branch lengths of the tree to maximize its likelihood value. The task can be efficiently solved by numerical analyses such as Brent or Newton-Raphson's methods. Searching the maximum likelihood tree is an NP-Hard problem [28]. Numerous heuristic methods have been proposed to build maximum-likelihood trees for large datasets with thousands of sequences.

Maximum likelihood methods that use the combination of stepwise addition tree construction algorithm and hill-climbing optimizations have been developed for searching maximum likelihood trees such as RAxML [10]. Similarly, Guidon and Gascuel proposed the PhyML method for building large maximum likelihood trees. The key idea of the PhyML method is a quick hill-climbing method based on NNI operations. The PhyML algorithm tries to perform multiple independent good swaps of nearest neighbor subtrees simultaneously instead of performing one swap per time. The strategy allows PhyML to handle datasets with thousands of sequences, and result in reasonably good maximum-likelihood trees.

The quartet-based approach is another strategy to build phylogenetic trees. For each quartet of four sequences, there are only three possible binary unrooted tree structures (or quartet trees, see Figure 4). Thus, it is easy to determine the maximum likelihood quartet tree. The quartet-based approach starts from a unique tree of three sequences, and iteratively adds new sequences into the current tree to build a whole tree. The key idea of quartet-based methods for inserting new sequences is the condition that if $T(a, b|c, y)$ is the best quartet tree of four sequences including a new sequence y and three leaves a, b, c of the current tree, the new sequence y should not be placed on the path connecting two leaves a and b. The quartet-based methods insert new sequences into the current tree such that the condition is hold for most of the available quartets.

Quartet puzzling method is the first quartet-based method to build a maximum likelihood tree [29]. The method evaluates all n^4 possible quartets to build a phylogenetic tree for n sequences. It is able to handle datasets with a few hundred sequences. To overcome the high complexity, the Important Quartet Puzzling (IQP) method has been proposed to build trees with thousands of sequences. The key idea of the IQP method is the definition of the so-called "important quartet" that consists of closely related sequences [9]. The IQP method selects only n^2 important quartets to build a phylogenetic tree. More importantly, the IQP algorithm is combined with the quick hill-climbing optimization based on the nearest neighbor interchange operations to further optimize the quartet-based constructed tree. Experiments on both real and simulated datasets showed that the combined method, IQPNNI, found better maximum likelihood trees than the PhyML method. Note that the IQPNNI method was implemented to run in parallel mode to handle large datasets [30].

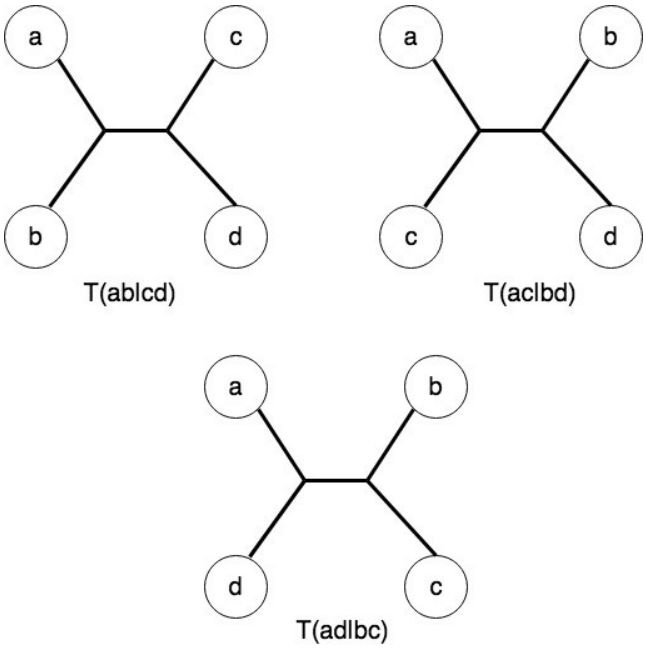

Figure 4. There different quartet trees topologies of a quartet.

The IQPNNI algorithm was improved into an efficient stochastic search algorithm for building large maximum likelihood trees [11]. They implemented IQ-TREE software that combined quick hill-climbing optimizations and a stochastic perturbation method to search maximum likelihood trees. The IQ-TREE perturbs current local optimal trees to escape from the current local optimal points and subsequently optimizes the perturbed trees by quick hill-climbing optimizations to search for the global optimal tree. The searching process is repeated several times and the best local optimal tree found will be considered as the best tree. Experiments showed that IQ-TREE was better than both

TABLE III
THE STRUCTURAL VARIANTS IN THE GENOMES.

	1	2	3	4	5	6	7	8	9
Human	G1	G2	G3	G4	G5	G6	G7	G8	G9
Gorilla	G1	G7	**G6**	**G5**	**G4**	**G3**	G7	G8	G9
Monkey	-	G5	G6	G7	G8	G9	**G2**	**G3**	**G4**
Dog	-	G5	G6	G7	G8	G9	**G4**	**G3**	**G2**

PhyML and RAxML methods in a majority number of cases tested. The IQ-TREE software is user-friendly and widely used by biologists for studying the evolution of species from molecular data.

IV. PHYLOGENOMIC ANALYSIS

Analyzing whole genomes to investigate the relationships among species is a comprehensive and challenging problem. In phylogenetic analysis, a genome is separated into a list of loci each corresponds to a gene or a region of interest in the genome. Homologous loci are aligned to create multiple sequence alignments. As a result, the input for phylogenomic analysis is a list of multiple sequence alignments. Phylogenetic tree construction approaches have been expanded to build phylogenomic trees from a list of multiple sequence alignments.

The first challenge in analyzing whole genomes is the occurrence of structural variants in the genomes. Besides variants occurring inside genes, other common types of structural changes are gene insertions/deletions, gene inversions (inverting the order of genes in the genome), gene transpositions (moving a number of genes in the genome from one position to another position in the genome), and inverted transpositions (combining both gene inversion and gene transposition events into one event). Table III demonstrates an example of structural variants in the genomes, *e.g.*, there is a gene inversion between the human genome and gorilla genome (genes G3, G4, G5, G6 in the human genome were inverted into G6, G5, G4, G3 in the gorilla genome).

The structural changes resulted in genomes with different structures. The structural difference between genomes can be used as phylogenetic signals for studying the evolution of species, *i.e.*, the number of structural changes to explain the structural difference between two genomes can be considered as the genetic distance between the genomes to evaluate their relationships. Overall, the genetic distance between two genomes is a combination of character changes inside genes and structural changes. Weighting and combining these changes properly for estimating the overall genetic distance between genomes enable us to build better phylogenetic trees using distance-based methods [31].

The second challenge when analyzing the genomes is model selection. The evolutionary models including rate models and substitution models might vary among loci. One model cannot properly reflex the evolutionary process of all loci. The phylogenomic analyses should use model selection methods to assign a proper model for each locus. As estimating model parameters is not strongly affected by the tree structure, model selection methods normally include two steps: building an initial tree and estimating model parameters based on the initial tree and the alignment. Building an initial tree can be done efficiently by distance-based methods such as NJ method. Provided that the initial tree is reasonably close to the best tree, it is good enough to estimate model parameters. Estimating model parameters for an alignment based on the initial tree can be solved by numerical optimization methods such as Brent's algorithm or more efficient Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm. Software like IQ-TREE provides us options to automatically determine the best model for each locus when analyzing whole-genome datasets. Note that normally we do not optimize parameters of amino acid models from each alignment because each alignment does not contain sufficient data to estimate a large number of parameters.

Finally, the computational expense is a critical burden of phylogenomic inferences. A genome dataset contains several dozens to thousands of genomes with the length up to several hundred million nucleotides. To overcome the problem, more efficient heuristic algorithms in terms of both running time and memory requirement should be developed to handle whole-genome datasets. As genomes are separated into loci, parallel computing is a promising approach to perform phylogenomic inference on individual alignments simultaneously. Most of the current widely-used phylogenetic software such as RAxML or IQ-TREE provide options to conduct phylogenetic inferences in parallel.

V. DISCUSSIONS

Phylogenetic inference is a core study in molecular biology. It is an active research field for several decades and the main focus of prominent researcher groups. Phylogenetic reconstruction for single or several genes perhaps come to age. The distance-based methods are able to build large reasonable phylogenies that could be used as starting points to search maximum parsimony or maximum likelihood trees. Nowadays, maximum likelihood methods such as RAxML, PhyML or IQ-TREE are able to efficiently construct trees with thousands of sequences.

All popular phylogenetic tree reconstruction software are based on heuristic search methods, therefore, the results from different software, or even from different runs of the same software, might not completely congruence. It is

especially true when analyzing datasets whose phylogenetic signals support polytomy tree structures. We recommend researchers to perform bootstrap analyses to assess the reliability of branches in the constructed tree. Although phylogenetic bootstrapping is computationally expensive, ultrafast bootstrap methods such as UFBoot2 [32] are able to build large bootstrap trees in an acceptable time.

Determining proper evolutionary models (*i.e.*, site rate models and/or substitution models) for datasets under the study is very critical in phylogenetic inferences. Using wrong evolutionary models in analyzing data will lead to inaccurate results [32]. The evolutionary models are normally selected from a list of existing models, and the model parameters can be directly estimated from the input data.

The advance of sequencing technologies has produced large genome datasets consisting of dozens to thousands of genomes with the length up to billion nucleotides. The large genome datasets provide us an unprecedented opportunity to study the relationships among species from their whole genomes. However, new efficient computational methods should be developed for phylogenomic inferences. The relationships among genomes should be analyzed at both levels: point level (*i.e.*, nucleotide/amino acid changes) and structural level (gene insertions/deletions as well as gene rearrangements). Combining changes at both levels to have a comprehensive evaluation is a new challenge for researchers in phylogenomic analyses. Another challenge in analyzing whole genomes is the heterogeneity of evolutionary processes between loci. Thus, determining proper evolutionary models is very critical when analyzing multiple genes or whole genomes. Currently, several software such as IQ-TREE are able to perform phylogenomic inferences for large genome datasets.

REFERENCES

[1] N. Saitou and M. Nei, "The neighbor-joining method: a new method for reconstructing phylogenetic trees." *Molecular Biology and Evolution*, vol. 4, no. 4, pp. 406–425, 1987.

[2] O. Gascuel, "BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data." *Molecular Biology and Evolution*, vol. 14, no. 7, pp. 685–695, 1997.

[3] L. S. Vinh and A. von Haeseler, "Shortest triplet clustering: reconstructing large phylogenies using representative sets," *BMC Bioinformatics*, vol. 6, no. 1, p. 92, 2005.

[4] J. C. Wilgenbusch and D. Swofford, "Inferring evolutionary trees with PAUP*," *Current Protocols in Bioinformatics*, no. 1, pp. 6–4, 2003.

[5] P. A. Goloboff, J. S. Farris, and K. C. Nixon, "TNT, a free program for phylogenetic analysis," *Cladistics*, vol. 24, no. 5, pp. 774–786, 2008.

[6] D. T. Hoang, L. S. Vinh, T. Flouri, A. Stamatakis, A. von Haeseler, and B. Q. Minh, "MPBoot: fast phylogenetic maximum parsimony tree inference and bootstrap approximation," *BMC Evolutionary Biology*, vol. 18, no. 1, p. 11, 2018.

[7] A. Varón, L. S. Vinh, and W. C. Wheeler, "POY version 4: phylogenetic analysis using dynamic homologies," *Cladistics*, vol. 26, no. 1, pp. 72–85, 2010.

[8] S. Guindon and O. Gascuel, "A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood," *Systematic Biology*, vol. 52, no. 5, pp. 696–704, 2003.

[9] L. S. Vinh and A. von Haeseler, "IQPNNI: moving fast through tree space and stopping in time," *Molecular Biology and Evolution*, vol. 21, no. 8, pp. 1565–1571, 2004.

[10] A. Stamatakis, "Using RAxML to infer phylogenies," *Current Protocols in Bioinformatics*, vol. 51, no. 1, pp. 6–14, 2015.

[11] L.-T. Nguyen, H. A. Schmidt, A. Von Haeseler, and B. Q. Minh, "IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies," *Molecular Biology and Evolution*, vol. 32, no. 1, pp. 268–274, 2015.

[12] J. Thompson, D. Higgins, and T. Gibson, "ClustalW," *Nucleic Acids Research*, vol. 22, no. 22, pp. 4673–4680, 1994.

[13] R. C. Edgar, "MUSCLE: multiple sequence alignment with high accuracy and high throughput," *Nucleic Acids Research*, vol. 32, no. 5, pp. 1792–1797, 2004.

[14] F. Joseph, *Inferring Phytogenies*. Sunderland, MA, USA: Sinauer Associates, 2003.

[15] Z. Yang, "Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites." *Molecular Biology and Evolution*, vol. 10, no. 6, pp. 1396–1401, 1993.

[16] S. Whelan and N. Goldman, "A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach," *Molecular Biology and Evolution*, vol. 18, no. 5, pp. 691–699, 2001.

[17] S. Q. Le and O. Gascuel, "An improved general amino acid replacement matrix," *Molecular Biology and Evolution*, vol. 25, no. 7, pp. 1307–1320, 2008.

[18] C. C. Dang, Q. S. Le, O. Gascuel, and V. S. Le, "FLU, an amino acid substitution model for influenza proteins," *BMC Evolutionary Biology*, vol. 10, no. 1, p. 99, 2010.

[19] D. T. Jones, W. R. Taylor, and J. M. Thornton, "The rapid generation of mutation data matrices from protein sequences," *Computational Applied Bioinformatics*, vol. 8, no. 3, pp. 275–282, 1992.

[20] L. L. Cavalli-Sforza and A. W. Edwards, "Phylogenetic analysis: models and estimation procedures," *American Journal of Human Genetics*, vol. 19, pp. 233–257, 1967.

[21] A. Rzhetsky and M. Nei, "Theoretical foundation of the minimum-evolution method of phylogenetic inference." *Molecular Biology and Evolution*, vol. 10, no. 5, pp. 1073–1095, 1993.

[22] W. H. Day and D. Sankoff, "Computational complexity of inferring phylogenies by compatibility," *Systematic Biology*, vol. 35, no. 2, pp. 224–229, 1986.

[23] O. Gascuel, "BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data." *Molecular Biology and Evolution*, vol. 14, no. 7, pp. 685–695, 1997.

[24] A. W. Edwards and Cavalli-Sforza, "The Reconstruction of Evolution," *Annals of Human Genetics*, vol. 27, pp. 105–106, 1963.

[25] W. M. Fitch, "Toward defining the course of evolution: minimum change for a specific tree topology," *Systematic Biology*, vol. 20, no. 4, pp. 406–416, 1971.

[26] R. Graham and L. Foulds, "Unlikelihood that minimal phylogenies for a realistic biological study can be constructed in reasonable computational time," *Mathematical Biosciences*, vol. 60, no. 2, pp. 133–142, 1982.

[27] Z. Yang, "Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate

methods," *Journal of Molecular Evolution*, vol. 39, no. 3, pp. 306–314, 1994.

[28] B. Chor and T. Tuller, "Maximum likelihood of evolutionary trees is hard," in *Annual International Conference on Research in Computational Molecular Biology*, 2005, pp. 296–310.

[29] K. Strimmer and A. Von Haeseler, "Quartet puzzling: a quartet maximum-likelihood method for reconstructing tree topologies," *Molecular Biology and Evolution*, vol. 13, no. 7, pp. 964–969, 1996.

[30] B. Q. Minh, L. S. Vinh, A. Von Haeseler, and H. A. Schmidt, "PIQPNNI: parallel reconstruction of large maximum likelihood phylogenies," *Bioinformatics*, vol. 21, no. 19, pp. 3794–3796, 2005.

[31] L. S. Vinh, A. Varón, and W. C. Wheeler, "Pairwise alignment with rearrangements," *Genome Informatics*, vol. 17, no. 2, pp. 141–151, 2006.

[32] D. T. Hoang, O. Chernomor, A. Von Haeseler, B. Q. Minh, and L. S. Vinh, "UFBoot2: improving the ultrafast bootstrap approximation," *Molecular Biology and Evolution*, vol. 35, no. 2, pp. 518–522, 2018.

Le Sy Vinh obtained PhD in Bioinformatics from Heinrich Heine University, Duesseldorf, Germany 2005, subsequently followed a postdoc fellowship at American Museum of Natural History, NYC from 2005 to 2008. He is currently the Dean of the Faculty of Information Technology, University of Engineering and Technology, Vietnam National University, Hanoi.

Le Sy Vinh is an expert in phylogenetic analysis, the author of widely-used software such as IQPNNI, POY4, UFBoot2. He is the group leader of many human genome projects in Vietnam including the first Vietnamese human genome, building the comprehensive Vietnamese human genome database, or Autism spectrum disorder in Vietnamese children.