The Author(s) *BMC Systems Biology* 2017, **11**(Suppl 4):79
DOI 10.1186/s12918-017-0453-x

BMC Systems Biology

RESEARCH Open Access

# Improvement of phylogenetic method to analyze compositional heterogeneity

Zehua Zhang[1], Kecheng Guo[1], Gaofeng Pan[1], Jijun Tang[1,2] and Fei Guo[1*]

*From* The 10th International Conference on Systems Biology (ISB 2016)
Weihai, China.19-22 August 2016

## Abstract

**Background:** Phylogenetic analysis is a key way to understand current research in the biological processes and detect theory in evolution of natural selection. The evolutionary relationship between species is generally reflected in the form of phylogenetic trees. Many methods for constructing phylogenetic trees, are based on the optimization criteria. We extract the biological data via modeling features, and then compare these characteristics to study the biological evolution between species.

**Results:** Here, we use maximum likelihood and Bayesian inference method to establish phylogenetic trees; multi-chain Markov chain Monte Carlo sampling method can be used to select optimal phylogenetic tree, resolving local optimum problem. The correlation model of phylogenetic analysis assumes that phylogenetic trees are built on homogeneous data, however there exists a large deviation in the presence of heterogeneous data. We use conscious detection to solve compositional heterogeneity. Our method is evaluated on two sets of experimental data, a group of bacterial 16S ribosomal RNA gene data, and a group of genetic data with five homologous species.

**Conclusions:** Our method can obtain accurate phylogenetic trees on the homologous data, and also detect the compositional heterogeneity of experimental data. We provide an efficient method to enhance the accuracy of generated phylogenetic tree.

**Keywords:** Phylogenetic analysis, Bayesian inference, Multi-chain Markov chain Monte Carlo, Conscious detection, Compositional heterogeneity

## Background

Phylogenetic analysis keeps an important role to understand current research in the biological processes and detect theory in evolution of natural selection. We extract the biological data via modeling features, and then compare these characteristics to study the biological evolution between species. The evolutionary relationship between species is generally reflected in the form of phylogenetic trees. Phylogenetic analysis can help to understand the evolutionary history of biological process, and become important data source for the development of large scale genomic data [1].

Many methods for constructing the phylogenetic tree, are based on optimization criteria, such as maximum parsimony, maximum likelihood and minimum evolution. Maximum parsimony (MP) approach [2, 3] examines all possible topologies or a certain number of topologies, which are likely to choose real phylogenetic tree or approximate phylogenetic tree with fewest evolutionary changes. Maximum likelihood (ML) approach [4, 5] tries to estimate trees by formulating a probabilistic model of evolution and applying known statistical method. It involves that phylogenetic tree yields the highest probability of evolutionary relationship. Minimum evolution (ME) approach [6] searches for the phylogenetic tree that minimizes total branch lengths. It is based on the assumption

*Correspondence: fguo@tju.edu.cn
[1]School of Computer Science and Technology, Tianjin University, 92 Weijin Road, Nankai District, Tianjin, People's Republic of China
Full list of author information is available at the end of the article

that the phylogenetic tree with smallest branch lengths is most likely to be the true one.

The correlation model of phylogenetic analysis assumes that phylogenetic trees are built on homogeneous data [7–10]. However, there exists a large deviation in the presence of heterogeneous data. As early as twenty years ago, there is first computational method [11] to detect heterogeneity problem, which makes people to doubt the credibility of phylogenetic analysis. Later, Markov model [12] of DNA sequence is used in the system development. Jukes-Cantor model [13] has been improved and taken into account unequal nucleotide compositions, different rates of changes from one nucleotide to another, variations in the form of invariant sites, and discrete gamma-distributed rates of variable sites. At the same time, researchers realize that the process of evolution would be different because of various evolutionary trees. It is obvious that the global rate can be often observed in fast and slow evolutionary species.

In this paper, we use maximum likelihood and Bayesian inference method to establish phylogenetic trees; multi-chain Markov chain Monte Carlo sampling method can be used to select optimal phylogenetic tree, resolving local optimum problem. We use two different instantaneous rate matrices, which is symmetrical and implies time-reversibility. We allow more than one composition vector to model compositional heterogeneity, because the overall model is tree-heterogeneous. The analysis is not reversible, and the likelihood depends the position of root. Compared to bootstrapping, Markov chain Monte Carlo yields a much larger sample of trees in the same computational time.

The correlation model of phylogenetic analysis assumes that phylogenetic trees are built on homogeneous data, however there exists a large deviation in the presence of heterogeneous data. The sample of trees produced by Markov chain Monte Carlo is highly auto-correlated, whereas many fewer bootstrapping replicates are sufficient. We make a conscious detection of phylogenetic tree produced by multi-chain Markov chain Monte Carlo sampling, analyzing multiple sampling and comparing different samples obtained from estimated values. We use conscious detection to solve compositional heterogeneity. Our method is evaluated on two sets of experimental data, a group of bacterial 16S ribosomal RNA gene data, and a group of genetic data with five homologous species. Our method can obtain accurate phylogenetic tree on the homologous data, and also detect the compositional heterogeneity of experimental data. We provide an efficient method to enhance the accuracy of generated phylogenetic tree.

## Method

We construct a phylogenetic tree for a set of DNA sequences. Our method generally contains following processes: aligning sequence [14–16], building phylogenetic trees, and selecting phylogenetic tree.

### Aligning sequence

The genetic information storage location has some differences on distinct species, such as information length and carrier of genetic information. These differences will affect our subsequent analysis. Therefore, we should arrange all possible similar sites in the same position, via a progressive algorithm of multiple sequence alignment. We adopt representational evolutionary multiple sequence alignment algorithm, called ClustalW [17–19]. It displays the alignment score, in form of identities, similarities and differences, and a guide tree of evolutionary relationship between aligned sequences.

### Building phylogenetic trees

The phylogenetic tree consists of many nodes and branches, where the node represents a taxon, namely species or sequence; the branch represents the evolutionary relationship between species [20, 21]. All nodes are divided into external nodes and internal nodes. In general, the external node represents actual observed taxon, the internal node represents location of evolutionary event.

#### *Phylogeny model*

Given the genetic information, we need the specific phylogeny model to predict evolutionary tree. First, we use the substitution model in terms of conversion rate. In general, the instantaneous conversion matrix is expressed as follows.

$$Q = \begin{pmatrix} -\mu(a\pi_C + b\pi_G + c\pi_T) & \mu a\pi_C & \mu b\pi_G & \mu c\pi_T \\ \mu g\pi_A & -\mu(g\pi_A + d\pi_G + e\pi_T) & \mu d\pi_G & \mu e\pi_T \\ \mu h\pi_A & \mu j\pi_C & -\mu(h\pi_A + j\pi_C + f\pi_T) & \mu f\pi_T \\ \mu i\pi_A & \mu k\pi_C & \mu l\pi_G & -\mu(i\pi_A + k\pi_C + l\pi_G) \end{pmatrix}$$