Theoretical Computer Science 524 (2014) 48–58

Contents lists available at ScienceDirect

Theoretical Computer Science

www.elsevier.com/locate/tcs

Is the Protein Model Assignment problem under linked branch lengths NP-hard?

Kassian Kobert [a,*], Jörg Hauser [a], Alexandros Stamatakis [a,b]

[a] Heidelberg Institute for Theoretical Studies, Germany
[b] Karlsruhe Institute of Technology, Institute for Theoretical Informatics, Postfach 6980, 76128 Karlsruhe, Germany

ARTICLE INFO

Article history:
Received 4 October 2012
Received in revised form 14 October 2013
Accepted 28 December 2013
Communicated by R. Giancarlo

Keywords:
NP-hard
Model selection
Phylogenetics
Maximum likelihood
Phylogenomics

ABSTRACT

In phylogenetics, computing the likelihood that a given tree generated the observed sequence data requires calculating the probability of the available data for a given tree (topology and branch lengths) under a statistical model of sequence evolution. Here, we focus on selecting an appropriate model for the data, which represents a generally non-trivial task. The data is represented as a so-called multiple sequence alignment. That is, each individual sequence of any one species (taxa) is arranged (aligned) in such a way, that the characters of all species at a given position (site) are assumed to share a common evolutionary history. It is well known, that an inappropriate model, which does not fit the data, can generate misleading tree topologies [3,4,26].

More specifically, we consider the case of partitioned protein sequence alignments. This means that the sites of the alignment may be clustered together into different partitions. Each partition may have an individual model of evolution. Our objective is to maximize the likelihood of the per-partition protein model assignments (e.g., JTT, WAG, etc.) when branches are linked across partitions on a given, fixed tree topology. That is, branch lengths are not estimated individually for each partition. Linked branch lengths across partitions substantially reduce the number of free parameters.

For p partitions and $|M|$ possible substitution models, there are $|M|^p$ possible model assignments. Since the number of combinations grows exponentially with p, an exhaustive search for the highest scoring assignment is computationally prohibitive for $|M| > 1$. We show that the problem of finding the optimal protein substitution model assignment under linked branch lengths on a given, tree topology, is NP-hard. Our results imply that one should employ heuristics to approximate the solution, instead of striving for the exact solution. Alternatively, the problem can be simplified by relaxing the assumptions.

1. Introduction and problem definition

1.1. Motivation and related work

In phylogenetics, many of the questions that we try to answer have been shown to be hard (NP-hard) to solve [1,8,14]. Among these are some of the most fundamental problems, such as finding the maximum likelihood tree for a given multiple sequence alignment [29,6] or even finding an optimal multiple sequence alignment [10], which are proven to be NP-hard. Some problems may not even have a unique solution, as is the case with finding the Maximum-Likelihood phylogeny [33].

* Corresponding author.

http://dx.doi.org/10.1016/j.tcs.2013.12.022

Here we are not interested in the actual phylogenetic tree search, but in the optimal assignment of evolutionary models to partitions of a partitioned multiple sequence alignment. At present, a plethora of empirical protein substitution models is available, such as WAG, JTT, DAYHOFF, etc. [35,20,23] some of which are collections of substitution matrices that contain different matrices such as the PAM or BLOSUM families [9,19]. It is well known, that a model, which does not fit the data well, can produce misleading tree topologies [3,4,26]. The models are provided in the form of an instantaneous 20×20 substitution matrix and the base frequencies (prior probabilities) of the states. Given this matrix (usually denoted as Q*-matrix*), one can calculate the transition probabilities from one state to another for a given time/branch length t. If each partition can be evaluated independently from the others, this task is almost trivial and an optimal solution can be found in polynomial time. However, if we assume that the branch lengths of the phylogenetic tree are jointly estimated over all partitions, the model choice for each partition is no longer independent from the choice of the models allocated to the other partitions. Under this assumption, the optimal assignment of models to partitions, with respect to the phylogenetic likelihood, is NP-hard, even if we assume a fixed tree topology.

When analyzing large multi-gene datasets joint branch length estimates can be used to reduce the number of free model parameters and thereby avoid over-parameterizing the model. Each set of independent per-partition branch lengths increases the number of model parameters by $2n - 3$ where n is the number of taxa. Therefore, the option to link branch lengths is offered in numerous phylogenetic tools such as RAxML [32] and PartitionFinder [24]. Numerous analyses on multi-gene alignments make use of this feature (e.g., [16,25,31]). Other results suggest that branch lengths may, under certain conditions, inherently be correlated across partitions [21], which provides an additional motivation to link branch lengths across partitions.

Tests on real-world data-sets performed by Hauser et al. [18] revealed that suboptimal model assignments under linked branch lengths can change the final tree topologies. They carried out tests on two previously published multi-gene data-sets [27,37] using RAxML-Light version 1.0.5 [32]. On these datasets, a total of 150 runs were conducted, on randomly chosen subsets containing three partitions and 50 species each. Thereafter, the best model assignment (with respect to its log likelihood score on the same fixed tree) was determined for each subset using linked *and* unlinked branch lengths. In 57% of the cases these model assignments were not identical. For the cases (subsets) where the model assignments differed, tree searches with RAxML under linked branch lengths using the two alternative model assignments were conducted. For 86% of these runs, the inferred best-known maximum likelihood trees are different. On average, the Robinson–Foulds distance [28] between the different trees inferred under the optimal and suboptimal model amounted to 9%. In other words, using the optimal protein model assignment under linked branch lengths on empirical data frequently yields a different tree topology with respect to the tree obtained from a suboptimal model assignment. Thus, the Protein Model Assignment problem (*PMA**) 'matters' since it alters the inferred tree topology. All data-sets from Hauser et al. are available for download via http://exelixis-lab.org/material/pma.tar.gz.

1.2. Protein Model Assignment problem

We define the Protein Model Assignment problem (*PMA**) as follows: Find the best-fit model from a set of available models for each partition of a protein alignment on some given, fixed, tree topology. Further assume that the branch lengths are linked across partitions. In other words, the branch lengths are estimated/optimized jointly across all partitions of the alignment. The following is a more formal definition:

Let M be a set of evolutionary models. Usually a model is defined by its Q*-matrix*. Here, the evolutionary models from which the Protein Model Assignment problem (*PMA**) can choose, are regarded as probability functions whose values represent the transition probability from one state to another, given a certain amount of time t, and the equilibrium frequencies for each state. The matrix and the frequencies are required for the actual likelihood calculations. We introduce this abstract view to avoid the calculations required for obtaining the transition-probabilities from the instantaneous transition rates in Q.

We denote a given model M_i with k states as:

$$M_i = (P, \Pi), \quad \text{where } \Pi \in [0,1]^k, \tag{1}$$

$$P : \mathbb{R} \rightarrow [0,1]^{k \times k}. \tag{2}$$

Here $P_{X,Y}(t) := P(X \rightarrow Y|t)$ is the probability of a transition/mutation from state X to state Y in time t and π_X is the equilibrium frequency of state X. For amino acid sequences we have 20 states, that is, $k = 20$.

Let S be an alignment for a set of taxa, divided into p partitions. That is, we define p partitions, $S_1, S_2, \ldots, S_p$ and each site s of S must satisfy $s \in S_i$ for some $i \in \{1, 2, \ldots, p\}$. Let $(T, \beta) = ((V, E), \beta(m))$ be a phylogenetic tree with nodes V, edges E and edge weights (branch lengths) β. The node set can be written as $V = N \cup I$, where N is the set of taxa (species) and I the set of inner nodes. The edge set $E \in V \times V$ must be such that the common tree properties are fulfilled and no edge $e \in E$ may satisfy $e \in N \times N$. The branch lengths $\beta(m)$ are given as edge weights under a chosen phylogenetic model assignment m. Formally we can write $\beta : M^p \rightarrow \mathbb{R}^{|E|}$.

Finding the optimal branch length configuration for a fixed tree topology and a given evolutionary model already represents a non-trivial numerical problem [13] and the solution may not be unique [5]. On real data, good approximations of the optimal branch length assignment can be computed efficiently, for example using the Newton–Raphson procedure [13,15].