Abstract

We consider the problem of reconstructing near-perfect phylogenetic trees using binary character states (referred to as BNPP). A perfect phylogeny assumes that every character mutates at most once in the evolutionary tree, yielding an algorithm for binary character states that is computationally efficient but not robust to imperfections in real data. A near-perfect phylogeny relaxes the perfect phylogeny assumption by allowing at most a constant number of additional mutations. We develop two algorithms for constructing optimal near-perfect phylogenies and provide empirical evidence of their performance. The first simple algorithm is fixed parameter tractable when the number of additional mutations and the number of characters that share four gametes with some other character are constants. The second, more involved algorithm for the problem is fixed parameter tractable when only the number of additional mutations is fixed. We have implemented both algorithms and shown them to be extremely efficient in practice on biologically significant data sets. This work proves the BNPP problem fixed parameter tractable and provides the first practical phylogenetic tree reconstruction algorithms that find guaranteed optimal solutions while being easily implemented and computationally feasible for data sets of biologically meaningful size and complexity.

Index Terms

computations on discrete structures, trees, biology and genetics

I. INTRODUCTION

Reconstruction of evolutionary trees is a classical computational biology problem [15], [24]. In the maximum parsimony (MP) model of this problem one seeks the smallest tree to explain a set of observed organisms. Parsimony is a particularly appropriate metric for trees representing short time scales, which makes it a good choice for inferring evolutionary relationships among individuals within a single species or a few closely related species. The intraspecific phylogeny problem has become especially important in studies of human genetics now that large-scale genotyping and the availability of complete human genome sequences have made it possible to identify millions of single nucleotide polymorphisms (SNPs) [26], sites at which a single DNA base takes on two common variants.

Minimizing the length of a phylogeny is the problem of finding the most parsimonious tree, a well known NP-complete problem [12]. Researchers have thus focused on either sophisticated heuristics or solving optimally for special cases (e.g. fixed parameter variants [1], [8], [20]).